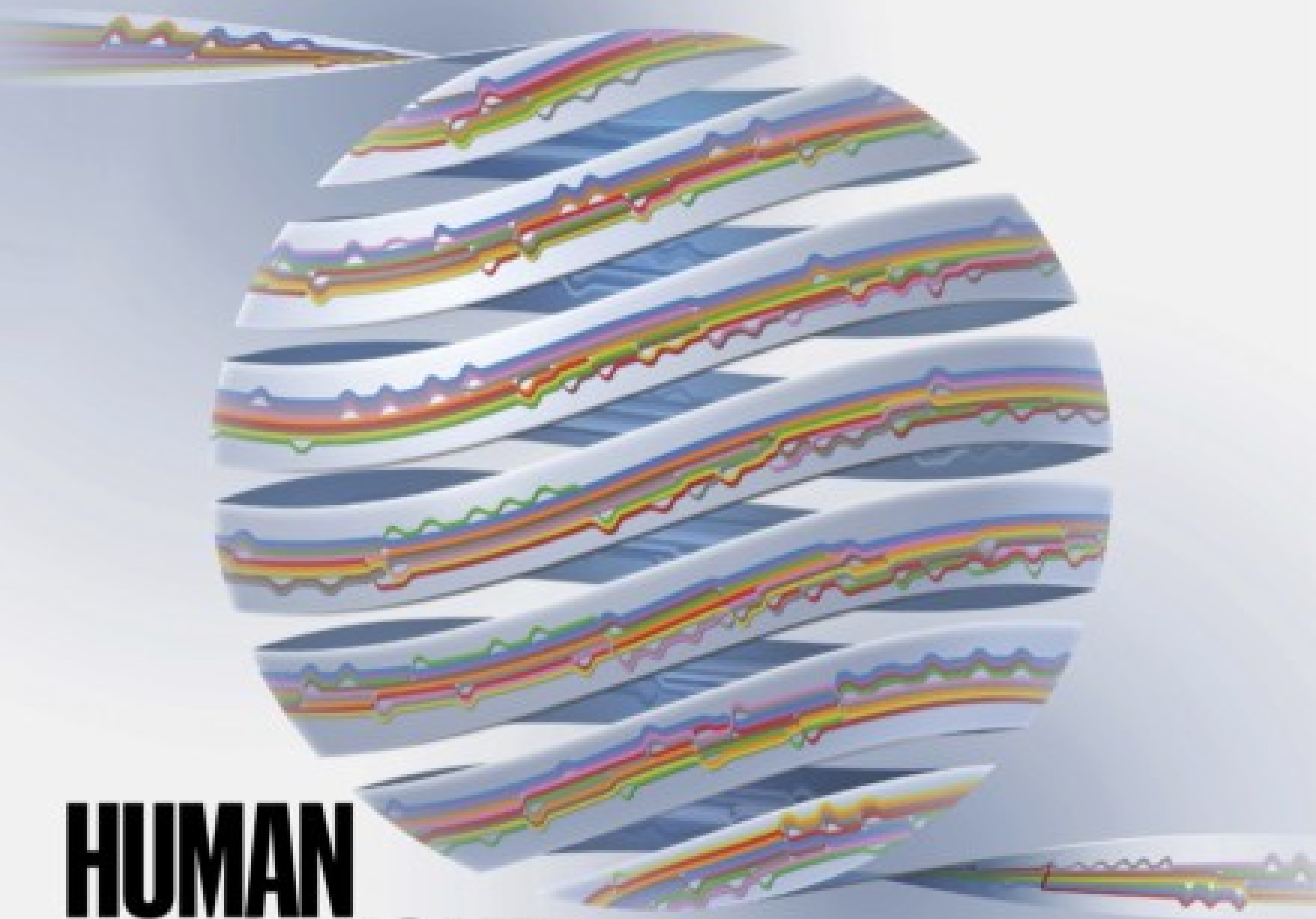
The international journal of science / 11 May 2023

# nature



# HUMAN PANGENOME

Data from 47 individuals combine to create reference resource that reflects human diversity

**Considered position**
The debate over priorities for autism research

**Predictive power**
Four steps to fashion a science of human–algorithm behaviour

**Screen test**
MicroLEDs induced to self-assemble into next-generation display

VISIT…

LANZAROTE Caliente.com

# Nature.2023.05.13

[Sat, 13 May 2023]


- **This Week**
- **News in Focus**
- **Books & Arts**
- **Opinion**
- **Work**
- **Research**
- **Amendments & Corrections**

# This Week

- **In space, failure is an option — often the only one** [09 May 2023]
  • Space companies should not lose heart when things go wrong. The first Moon missions failed repeatedly — and provided lessons on how to achieve success in space and beyond.
- **Why the fate of Turkish higher education hinges on the country's general election** [09 May 2023]
  • Abolishing the council that controls the universities would allow higher education to finally blossom.

- EDITORIAL
- 09 May 2023

# In space, failure is an option — often the only one

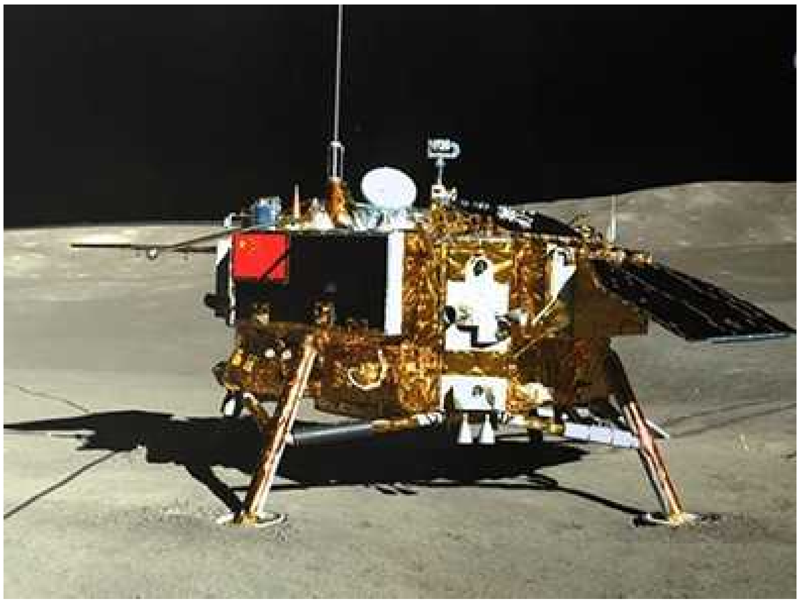
Space companies should not lose heart when things go wrong. The first Moon missions failed repeatedly — and provided lessons on how to achieve success in space and beyond.




The actor Ed Harris plays NASA flight director Gene Kranz in the 1995 film *Apollo 13*, directed by Ron Howard.Credit: Landmark Media/Alamy

“Failure is not an option,” NASA’s legendary flight-operations director Gene Kranz is said to have remarked, as seen in the 1995 film *Apollo 13*. Actor Ed Harris portrayed Kranz as he guided his team to save a spacecraft that had run into trouble on the way to the Moon. In the movie, as in real life, the three astronauts on the Apollo 13 mission pulled off a spectacular fix and returned safely to Earth.

Not all space ventures have such a tidy ending. A 2019 attempt by Israeli company SpaceIL to land on the Moon crashed. On 20 April this year, a spectacular intentional detonation ended the first major test flight of Starship, the world's largest rocket, which SpaceX in Hawthorne, California, is building to carry humans back to the Moon and to Mars. The craft had spun out of control four minutes after lifting off its launch pad in Texas. Five days later, a robotic mission from the Japanese company ispace, based in Tokyo, tried and failed to land safely on the Moon.



Moon mission failure: why is it so hard to pull off a lunar landing?

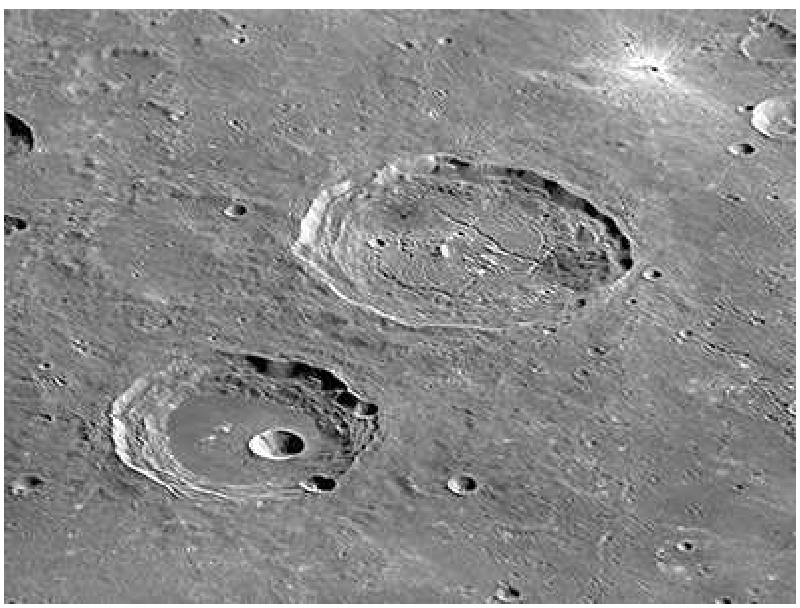
Its crash-landing was devastating not only for the small team of engineers that had built and flown the spacecraft, but also for others whose hopes were riding along with it. When the lander crashed, it pulverized two rovers designed to travel around and explore the Moon's surface — one built by the space agency of the United Arab Emirates and the other by the Japan Aerospace Exploration Agency. A camera system built by Canadensys Aerospace of Toronto, Canada, which was designed to photograph the rovers deploying, also went up in smoke.

The scientists and engineers involved should not be discouraged by these failures. Space is hard. This is a truism trotted out every time there's an attempt to launch from this planet or land on another. But it is accurate.

Those who wish to explore the cosmos should expect to fail — perhaps many times — before they can succeed.

Engineering requires iteration, time and again: first to design machines that might work and then to test them against as many possible scenarios in which things could, and of course do, go wrong.

As with all failures, the Starship explosion will teach scientists and engineers valuable lessons before the craft attempts to fly again. Along with studying why the rocket's 33 engines did not all ignite and burn as they were supposed to, SpaceX also needs to reckon with the extensive environmental damage caused when Starship's launch pelted nearby communities with more sand and debris than expected. NASA is watching this process closely, because it is depending on Starship for its plans to send humans back to the Moon in the coming years.



Private ispace Moon landing fails: researchers are investigating

Ispace, too, will study and learn from its mission's problems, ahead of a planned second launch next year. It looks as if the spacecraft ran out of propellant just before reaching the surface, causing it to crash-land.

Humans live in a culture in which it seems as if everything must go right the first time we try it, but that is not how successful products are developed,

nor how science unfolds — especially in space. Back in the 1960s, both the US space programme and the Soviet Union's equivalent went through a process of trial and error, as they tried time after time to land the first Moon missions, and failed repeatedly. Both learnt from each attempt and incorporated those lessons the next time around.

Kranz did not actually say "Failure is not an option" — although that didn't stop him using the phrase as the title of his 2000 autobiography. In research and development, failure is indeed an option. In fact, it can be a necessary learning objective on the path to success.

*Nature* **617**, 223 (2023)

*doi: https://doi.org/10.1038/d41586-023-01547-3*

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01547-3

- WORLD VIEW
- 09 May 2023

# Why the fate of Turkish higher education hinges on the country's general election



Abolishing the council that controls the universities would allow higher education to finally blossom.

- Taner Bilgiç

The Turkish general election on 14 May could be a watershed moment for our nation's higher education.

I entered university in 1983 as a student; since 1997, I have been a faculty member in industrial engineering at Boğaziçi University in Istanbul. In all that time, Turkish higher education has been under the iron fist of the central Council of Higher Education (YÖK). Imposed after a military coup

in 1980, when universities were seen as hotbeds of dissent, this body has far-reaching power over the academic and administrative matters of every university in Turkey, with few to no checks or balances.

In the 11 general elections since the restoration of civilian rule in 1983, political parties have pledged to abolish the YÖK. This includes the Justice and Development Party (AKP) of current president Recep Tayyip Erdoğan, which has been in power since 2002. Yet never in 21 years has the AKP attempted to follow through. Like other governments before it, it has instead used the YÖK as a convenient centralized tool for imposing its ideology.

After more than 40 years of this dysfunctional system, higher education in Turkey is in urgent need of reform.



What Turkey’s earthquake tells us about the science of seismic forecasting

The value of degrees obtained at many of our universities is debatable, and Turkey has the lowest employment rate of university-educated adults among countries in the Organisation for Economic Co-operation and Development. The YÖK is undoubtedly a big part of the problem. Universities have little or no authority to determine their administrative, financial and academic procedures. At public universities, salaries are fixed throughout the country, and tenure decisions are by central peer review.

These systems include perverse incentives. In 1993, Turkey began paying academics for each paper they published internationally. And since 2000, a scoring system dominated by international publications has been used to assess tenure claims. This has boosted the number of publications but has also increased instances of author–citation networks, dishonesty and predatory publishing.

Many talented young scientists have already left the country. More and more of my own students want to pursue their careers abroad.

But I see a real opportunity for change now, partly because science in Turkey is in the public eye as never before. The COVID-19 pandemic and February's devastating earthquakes in southeast Turkey have sent awareness of the value of science soaring. Scientists had been warning the public and policymakers for years about the consequences of such an earthquake and had made the scientific evidence readily available. It is clear to many that lives could have been saved had the science been heeded.

What's more, issues of academic freedom have become public talking points. At Boğaziçi University, the longest-running peaceful resistance of faculty members in Turkish academia has been ongoing since 2021, triggered by Erdoğan's appointment of a rector from outside the university with connections to the president's party. Peaceful student and faculty protests and a subsequent wave of disciplinary actions and student arrests caught the public's attention. In response to the protests, the YÖK doubled down, sacking three deans simultaneously, appointing its own replacements and unilaterally establishing schools of law and communications with no senate approval. The appointed rector changed the locks on the Istanbul Center for Mathematical Sciences to keep researchers out.



Turkey–Syria earthquake: what scientists know

But this was also something of a turning point. During these events, I was an elected member of the executive board of Boğaziçi University and chair of a senate subcommittee on restructuring university governance, set up in 2016. I was also among the faculty members who filed several lawsuits, some ongoing, against the university and the YÖK regarding the breaches of academic autonomy.

A June 2022 workshop grew out of these efforts, attended by some 45 academics from 13 universities. Our resulting report concluded that Turkish higher education must be rebuilt on a basis of academic freedom and university autonomy. The YÖK should be abolished and a new law should enshrine autonomous legal status for universities. Other principles include: secular, democratic, equitable and inclusive institutions; shared and bottom-up decision-making; accountability; and transparent budgeting.

Our report was sent to all political parties represented in parliament. The AKP's policy statement for the upcoming elections contains no reference to the YÖK. But the AKP is currently running neck and neck in the polls with the Nation Alliance, a coalition of six opposition parties. The alliance has adopted several of our recommendations, including abolishing the YÖK, as part of its platform. On 16 April, along with four colleagues from Boğaziçi University, I met with the alliance's presidential candidate, Kemal

Kılıçdaroğlu, who reaffirmed his support for acting on this plan if he assumes power after the elections.

Turkish academics, students and policymakers have the insight, experience and creativity to build a competitive, efficient, decentralized higher-education system. Regardless of who wins the elections, we must hold policymakers to their promises, guided by our fundamental values of academic freedom and university autonomy. Giving us an opportunity to do so will lift the chances of a whole nation.

*Nature* **617**, 225 (2023)

*doi: https://doi.org/10.1038/d41586-023-01548-2*

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01548-2

# News in Focus

- **COVID’s future: mini-waves rather than seasonal surges** [01 May 2023]
  • Three years after the start of the pandemic, SARS-CoV-2 shows no signs of settling into a seasonal pattern of spread, like influenza has.
- **Air pollution in China is falling — but there is a long way to go** [01 May 2023]
  • Easy gains from upgrading power-plant smokestacks will be strengthened only by deeper policy changes.
- **‘Armed groups entered the lab’: Sudan’s researchers flee violent military conflict** [28 Apr 2023]
  • Clashes have left hospitals and universities without water and power — and vulnerable to takeover by armed groups.
- **Alzheimer’s drug donanemab: what promising trial means for treatments** [04 May 2023]
  • Results suggest that the amyloid-targeting drug candidate slows cognitive decline in some people, but questions remain over its potential side effects.
- **Wuhan market samples fail to shed further light on COVID origins** [04 May 2023]
  • New analysis of genomic data from market swabs highlights their limitations.
- **‘A good day’: FDA approves world’s first RSV vaccine** [03 May 2023]
  • Green light from US regulatory agency for GSK’s jab caps off a decades-long quest.
- **How generative AI is building better antibodies** [04 May 2023]
  • Language models similar to those behind ChatGPT have been used to improve antibody therapies against COVID-19, Ebola and other viruses.
- **Mind-reading machines are here: is it time to worry?** [02 May 2023]
  • Neuroethicists are split on whether a study that uses brain scans and AI to decode imagined speech poses a threat to mental privacy.

- **‘The best way to get it right is to listen to us’ — autistic people argue for a stronger voice in research** [10 May 2023]
  • Despite broader acceptance of neurodiversity, autistic advocates and scientists are still fighting for a chance to set clinical and scientific priorities.
- **Human embryo science: can the world’s regulators keep pace?** [10 May 2023]
  • Concerned by the speed of research, policymakers are looking for alternative ways to establish rules on genome editing and other technologies.

---

| Next section | Main menu | Previous section |

- NEWS
- 01 May 2023
- Correction 04 May 2023

# COVID’s future: mini-waves rather than seasonal surges

Three years after the start of the pandemic, SARS-CoV-2 shows no signs of settling into a seasonal pattern of spread, like influenza has.

- Ewen Callaway




A woman is tested for COVID-19 in India, which has recorded a surge of infections driven by a SARS-CoV-2 variant called XBB.1.16.Credit: Narinder Nanu/AFP via Getty

Whether you call it a surge, a spike, a wave or perhaps just a wavelet, there are signs of a rise in SARS-CoV-2 infections — again. A growing proportion of tests in some countries are coming back positive, and new variants, most notably a lineage called XBB.1.16, are pushing aside older strains, fuelling some of the uptick in cases.

Welcome to the new normal: the 'wavelet' era. Scientists say that explosive, hospital-filling COVID-19 waves are unlikely to return. Instead, countries are starting to see frequent, less deadly waves, characterized by relatively high levels of mostly mild infections and sparked by the relentless churn of new variants.



Are repeat COVID infections dangerous? What the science says

Wavelets don't always create a dramatic spike in hospitalizations and deaths; their effects on health vary between countries. But the relentless series of wavelets looks very different from the slower, annual circulation patterns of influenza and cold-causing coronaviruses, and it seems increasingly unlikely that SARS-CoV-2 will settle into a flu-like rhythm anytime soon, say scientists.

"We haven't slowed down in the last year, and I don't see what factors would cause it do so at this point," says Trevor Bedford, an evolutionary biologist at the Fred Hutchinson Cancer Center in Seattle, Washington. "It will be a continually circulating respiratory disease. It may be less seasonal than things we're used to."

# Victorious variant

In March, scientists in India started seeing signs that a new SARS-CoV-2 variant was causing a rise in infections. The XBB.1.16 lineage has displaced others that drove case surges in India several months ago, says Rajesh Karyakarte, a microbiologist at Byramjee Jeejeebhoy Government Medical College in Pune, India. “We see that it has almost replaced all other variants in India, and we think the same thing will be followed everywhere.”

In a study[1] posted to the medRxiv.org preprint server, Karyakarte and his colleagues analysed more than 300 cases, from last December to early April, and found that XBB.1.16 infections tend to cause mild symptoms similar to those from earlier Omicron variants, with few hospitalizations and deaths. “We didn’t see much,” Karyakarte says. The study has not yet been peer reviewed.



How your first brush with COVID warps your immunity

The World Health Organization declared XBB.1.16 a ‘variant of interest’ on 17 April. But whether it or another new variant will cause a spike in infections in a particular country will probably depend on the size and timing of the country’s earlier waves, says Tom Wenseleers, an evolutionary biologist at the Catholic University of Leuven in Belgium.

He estimates that XBB.1.16 is spreading fairly rapidly in the United States, where it is now estimated to make up more than 11% of cases. In Europe,

the variant is less prevalent and is spreading more slowly. This could be due to Europe's relatively large and recent spike in infections caused by a closely related variant, XBB.1.5, which struck earlier in the United States.

## Wavelet, surge, repeat

Some countries are experiencing surges of infections three or four times each year, driven largely by the breakneck pace at which the virus continues to change, says Bedford. Currently, SARS-CoV-2's spike protein, in which most immunity-evading mutations occur, is evolving at twice the rate of a similar protein in seasonal influenza and about ten times as quickly as those of cold-causing 'seasonal' coronaviruses.

Influenza and common-cold coronaviruses cause seasonal epidemics in part because of favourable transmission conditions, such as people spending more time indoors during winter. The combination of rapid mutation and short-lived human immunity is probably preventing SARS-CoV-2 from settling into seasonal patterns of circulation, says Wenseleers.

The stubbornly high frequency of SARS-CoV-2 surges translates to large numbers of infections. Data from the now-defunct UK survey of SARS-CoV-2 prevalence suggested that the country experienced as many infections as it has residents in the last year, which equates to a 100% annual 'attack rate', says Bedford. In the future, "we can still imagine 50% attack rates every year, half the population getting infected", compared with around 20% with influenza.

## Emulating influenza

There is little doubt, however, that SARS-CoV-2's continuing ebb and flow is causing fewer problems than in the past.

In South Africa, the country's health systems would quickly send notice if COVID-19 hospitalizations and deaths were on the rise, says Waasila Jassat, a public-health specialist at the country's National Institute for Communicable Diseases in Johannesburg. "That doesn't seem to be have been the case for many, many months," she says.

In the year and half since Omicron emerged, COVID-19 deaths has remained stubbornly high and the toll has been around ten times higher than that typically caused by influenza, says Wenseleers. But, still, large infection waves are causing smaller ripples in hospitalizations and deaths. “It gives most people the hope that, in the coming years, the net toll of COVID will get comparable to influenza,” he says.

*Nature* **617**, 229-230 (2023)

*doi: https://doi.org/10.1038/d41586-023-01437-8*

### Updates & Corrections

- **Correction 04 May 2023**: An earlier version of this article misstated the timing of COVID-19 cases analysed in a new preprint.

## References

1. Karyakarte, R. P. *et al*. Preprint at medRxiv https://doi.org/10.1101/2023.04.22.23288965 (2023).

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01437-8

- NEWS
- 01 May 2023

# Air pollution in China is falling — but there is a long way to go

Easy gains from upgrading power-plant smokestacks will be strengthened only by deeper policy changes.

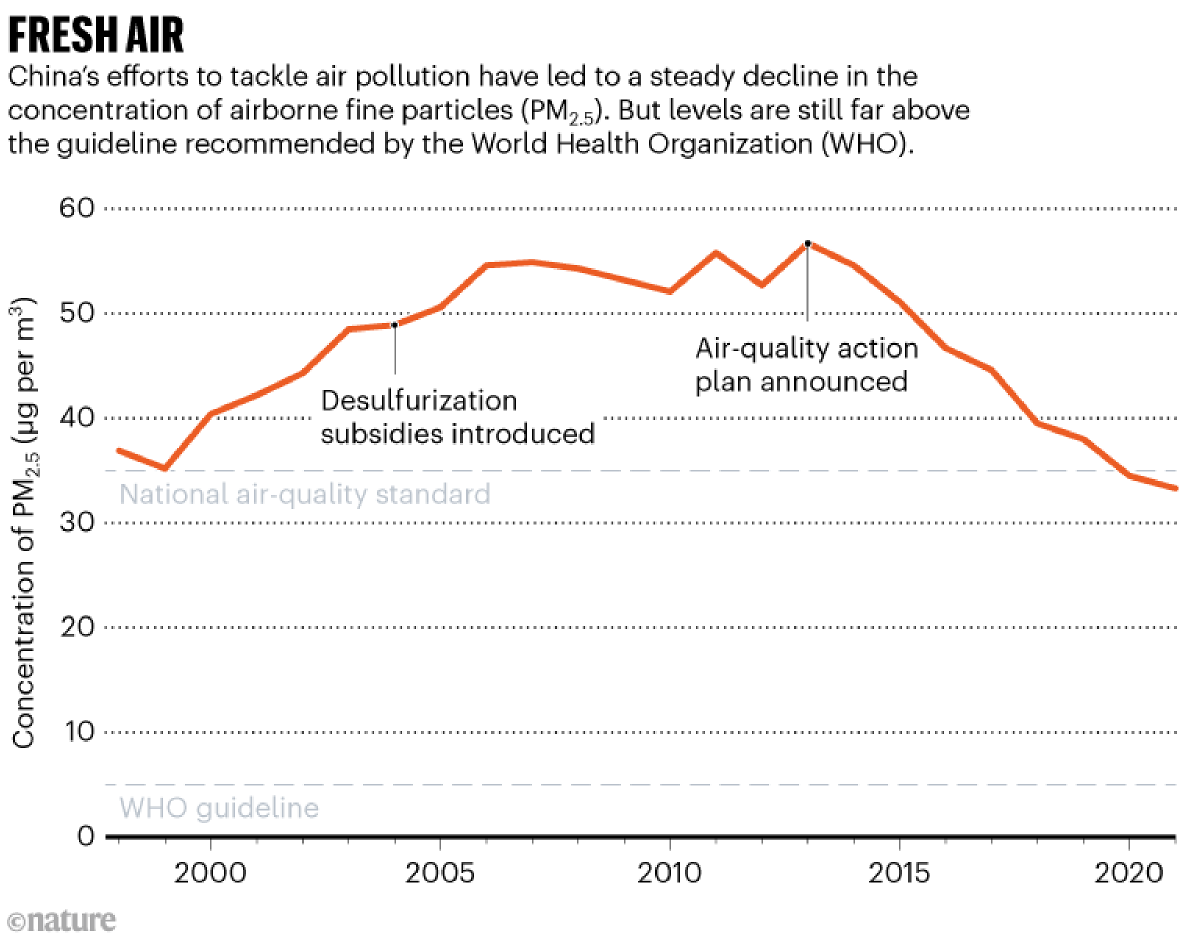
- Dyani Lewis



Parts of China continue to be affected by heavy smog.Credit: Getty

Over the past decade, China's once-pollution-choked skies have steadily improved, according to more than two decades of atmospheric measurements taken by NASA satellites. But researchers say that there is still a long way to go to clean China's air and protect the health of its citizens.

The speed at which China has reduced its air pollution has been "impressive", says Chi Li, an atmospheric scientist at Washington University in St. Louis, Missouri, owing to technological solutions and ambitious policies.




Source: Atmospheric Composition Analysis Group, Washington Univ. St. Louis

Each year, air pollution is responsible for more than four million premature deaths globally — including an estimated one million in China — primarily from heart disease, lung cancer and respiratory illnesses. Fine particulate

matter with a diameter of 2.5 micrometres or less — referred to as $PM_{2.5}$ — is the most concerning air pollutant, says Li.

The university's Atmospheric Composition Analysis Group, which Li is part of, monitors various pollutants and estimates their global health impacts. The data for China show that from the late 1990s, average annual $PM_{2.5}$ exposure in the country rose from 35 to more than 50 micrograms per cubic metre, before levelling out around 2006 at between 50 and 60[1]. Since 2013, $PM_{2.5}$ levels have steadily declined, and in 2021, the average annual exposure was 33.3 micrograms per cubic metre (see 'Fresh air'). That's below the nation's air-quality standard of 35, but still much higher than the recommended level of 5 set by the World Health Organization (WHO), based in Geneva, Switzerland.

## Smokestack solutions

The decline in $PM_{2.5}$ is the result of targeted efforts by China over the past two decades to address poor air quality. Upgrades to coal-fired power plants have had the biggest effect so far, says Qiang Zhang, an atmospheric scientist at Tsinghua University in Beijing.

Since 2004, the Chinese government has provided subsidies to retrofit smokestacks in coal-fired power plants with filters and other equipment to remove sulfur dioxide — a molecule that reacts with other compounds in the atmosphere to form $PM_{2.5}$ particles — from emissions.



How air pollution causes lung cancer — without harming DNA

In 2013, China released its air-pollution prevention and control action plan, which further tightened standards for industrial emissions, and shut down small, inefficient power generators and industrial operators.

An analysis by Zhang and his colleagues shows that these measures accounted for 81% of the reductions to $PM_{2.5}$ emissions between 2013 and 2017[2].

Further reductions could lead to fewer heavy-pollution days — driven by industrial emissions and cold weather that prevents dispersal — when the daily $PM_{2.5}$ concentration can exceed 200 micrograms per cubic metre. Last year, China's government set a target to eliminate heavy-pollution days by 2025.

But there's still a long way to go. In 2021, the WHO lowered its recommended annual-exposure limit for $PM_{2.5}$ from 10 to 5 micrograms per cubic metre, a level that most countries, including the United Kingdom, Germany, the United States and Canada, exceed.

## Energy transition

Zhang says that the air-quality improvements from end-of-pipe technologies, such as smokestack filters, that target industrial emissions will eventually be exhausted. “Energy and the climate policies would definitely play more important roles in the future,” he says. These include efforts to supply more households with natural gas or electric heating systems in parts of rural China that still rely on coal and wood-fired stoves to heat their homes.

Electrical grids in rural areas are being upgraded to accommodate the increased capacity required for domestic heating, says Zhang, and the renewable-energy sector is expanding. But “it’s still a long way to go” before coal-fired power is replaced, he says.

China’s goal to become carbon neutral by 2060 will help to achieve that transition, and will keep its air pollution trending downwards, says Li. “It will ultimately be more and more clean in the future if the electricity source gets cleaner as well,” he says.

*Nature* **617**, 230-231 (2023)

*doi: https://doi.org/10.1038/d41586-023-01452-9*

# References

1. van Donkelaar, A. *et al. Environ. Sci. Technol.* **55**, 15287–15300 (2021).
2. Zhang, Q. *et al. Proc. Natl Acad. Sci. USA* **116**, 24463–24469 (2019).

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01452-9

- NEWS
- 28 April 2023

# ‘Armed groups entered the lab’: Sudan’s researchers flee violent military conflict

Clashes have left hospitals and universities without water and power — and vulnerable to takeover by armed groups.

- Miryam Naddaf



Smoke rises during clashes between the Sudanese Armed Forces and the paramilitary Rapid Support Forces in Khartoum on 19 April.Credit: Omer Erdem/Anadolu Agency/Getty

Sudan's academics and students have been forced to abandon universities and residential campuses because of heavy artillery and aerial bombardment in the capital, Khartoum, and other major cities, as clashes have erupted between warring military factions.

The violence, which began on 15 April, is between members of the country's army, the Sudanese Armed Forces (SAF), and a rival paramilitary group called the Rapid Support Forces (RSF). Although a temporary ceasefire is in place, street battles and explosions are still being reported in Khartoum.

The conflict has killed more than 450 people and injured more than 4,000 others, according to the United Nations. Thousands more are on the move, looking to find a safe place. Most of the injured people are unable to access medical attention. Nearly two-thirds of hospitals in Khartoum are closed, says the World Health Organization.

"This was an unexpected conflict," says Deen Sharp, an urban geographer at the London School of Economics and Political Science, who is based in Khartoum. Sharp is compiling a list of academics in Sudan who need money, food, water and medicines. "No one was really prepared."

Sudan's people have lived mostly under military rule since they gained independence from the United Kingdom and Egypt in 1956. Mass protests in 2019 overthrew a three-decade dictatorship, but the army remains the most powerful institution.

*Nature* spoke to three researchers in Sudan who are among the millions of people caught up in the turmoil.



Hisham Bilal is a researcher in the department of sociology and anthropology at the University of Khartoum.Credit: Courtesy of Hisham Bilal

## HISHAM BILAL: ‘We didn’t have water’

*Anthropologist at the University of Khartoum.*

I woke up on the morning of 15 April to a phone call from my sister. I was planning to go to the university that day, but she told me she had heard the sound of clashes.

The University of Khartoum is close to both the SAF headquarters and the RSF headquarters. Almost 90 students and faculty members living on the university campus were stranded. They moved to the basement where they hid with some staff members for days. They were freed three days later, but one of the students died from gunfire and had to be buried on the grounds of the engineering campus.

I was living on another university campus with my wife and children, close to a military base. The electricity went off and we didn’t have water. Nobody was able to leave the campus to go anywhere to get any supplies, including food.

My apartment was hit by many bullets. I counted four and they entered from the balcony and the windows. At that point, we decided to leave no matter what might happen.



Shahinaz Bedri, director-general of the National Public Health Laboratory in Khartoum.Credit: Nayef Kheyri

## SHAHINAZ BEDRI: ‘Armed groups are entering health facilities’

*Pathologist and director-general of the National Public Health Laboratory in Khartoum.*

I left Khartoum yesterday with my mother, who has dementia and is bedridden and whose health has deteriorated. We headed to Port Sudan, 800 kilometres away from Khartoum by the Red Sea, because it’s safer and my mother can have access to health services.

I run the National Public Health Laboratory in Khartoum, which is the national reference regulatory laboratory for infectious diseases, with more than 500 staff members. On site, we have samples of COVID-19, influenza, measles, polio, tuberculosis and cholera, among others. We also have a

biosecure unit. Specimens are in a secure place, but we're worried that they might be mishandled by armed groups entering health facilities.

Several days ago, armed groups entered the lab, and ordered our security staff, who are trained to handle emergencies, to leave. We are worried that there might be a bio-risk for them if they open fridges, because they go in and they sabotage.

Power has been cut in several hospitals and in our lab. There are always electricity shortages in Sudan — that's why we have power from diesel generators. But now, because of this war, we can't refuel them.

By now, most of the unrefrigerated samples will have died. But we need to get people to come in and safely assess what's happening, deal with issues, and do bio- and chemical waste management.

The other thing is that we have very sensitive analytical chemistry equipment, such as gas-chromatography apparatus. Some are attached to gas cylinders like helium and hydrogen, so we don't want somebody going in and shooting. That's a worry for us.

## AZZA AHMED ABDEL AZIZ: 'The military is overturning democracies'

*Social and medical anthropologist at the Centre for Economic, Legal and Social Studies in Khartoum.*

The situation at the University of Khartoum has left people traumatized, but this is a general trauma: the violence has left bodies strewn across different parts of the city.

Forced to stay at home, academics are turning to social media to write analyses of the situation. But there's also a lot of propaganda — unidentifiable people who are just writing and spreading misinformation.

This is a challenge for academics working on the ground to counter. We have a problem here with the military overturning democracies in the post-

colonial era. It’s always the Sudanese state not accepting alternative ideologies.

*Nature* **617**, 231-232 (2023)

*doi: https://doi.org/10.1038/d41586-023-01478-z*

Readers seeking to support Sudan’s academic community can contact Scholars at Risk, a charitable organization based in New York City that helps threatened scholars worldwide.

These interviews have been edited for length and clarity.

---

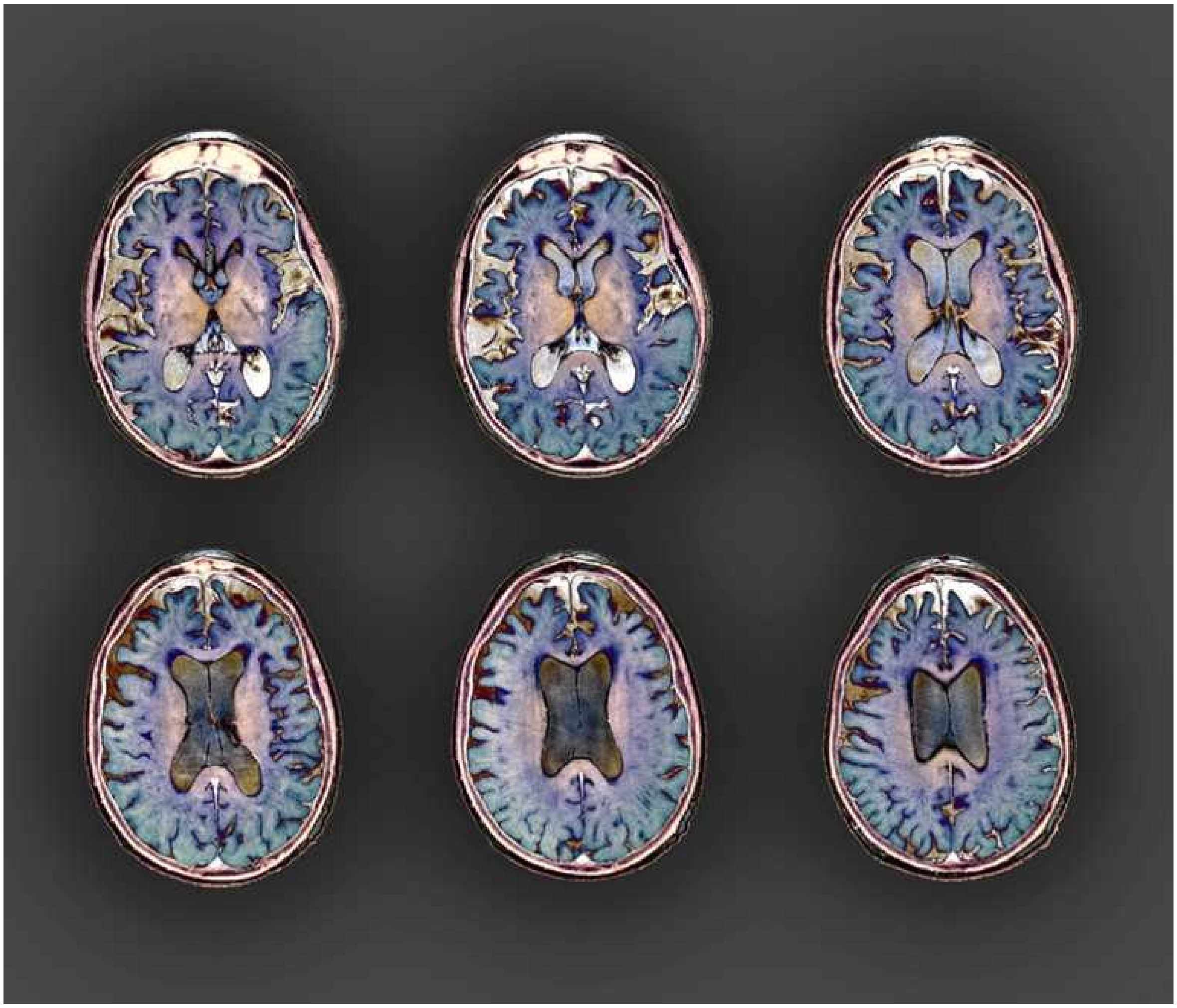
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01478-z

- NEWS
- 04 May 2023

# Alzheimer’s drug donanemab: what promising trial means for treatments

Results suggest that the amyloid-targeting drug candidate slows cognitive decline in some people, but questions remain over its potential side effects.

- Sara Reardon



An MRI scan shows shrinkage around the front and sides of the brain of a person with early-onset Alzheimer's disease.Credit: Zephyr/SPL

For the second time, an experimental drug has been shown to reduce the cognitive decline associated with Alzheimer’s disease. On 3 May, pharmaceutical company Eli Lilly announced in a press release that its monoclonal antibody donanemab slowed mental decline by 35% for some participants in a 1,736-person trial — a rate comparable to that for competitor drug lecanemab. But researchers warn that until the full results are published, questions remain as to the drug’s clinical usefulness, as well as whether the modest benefit outweighs the risk of harmful side effects.

Like lecanemab, donanemab targets amyloid protein, which is thought to cause dementia by accumulating in the brain and damaging neurons. The trial results provide strong evidence that amyloid is a key driver of

Alzheimer’s, says Jeffrey Cummings, a neuroscientist at the University of Nevada, Las Vegas. “These are transformative in an enormously important way from a scientific point of view,” he adds. “They’re terrific.”



FDA approves Alzheimer’s drug lecanemab amid safety concerns

But Marsel Mesulam, a neurologist at Northwestern University in Chicago, is more cautious. “The results that are described are extremely significant and impressive, but clinically their significance is doubtful,” he says, adding that the modest effect suggests that factors other than amyloid contribute to Alzheimer’s disease progression. “We’re heading to a new era — there’s room to cheer, but it’s an era that should make us all very sober, realizing that there will be no single magic bullet.”

In the press release, Eli Lilly said that people with mild Alzheimer’s who received donanemab showed 35% less clinical decline over 18 months than did those who received a placebo, and 40% less decline in their ability to perform daily tasks. The company, based in Indianapolis, Indiana, says that it will present the full results at a conference in July and publish them in a peer-reviewed journal. It plans to apply for approval by the US Food and Drug Administration (FDA) in the next two months.

## Promising treatments

FDA approval would make donanemab the third new Alzheimer's treatment in two years. In January, the agency granted accelerated approval to lecanemab, made by Biogen in Cambridge, Massachusetts, and Eisai in Tokyo. A study[1] published in November showed that lecanemab slowed cognitive decline in 1,800 patients by 27% over 18 months. The FDA had previously approved aducanumab, also made by Biogen and Eisai, on the basis of evidence that it could reduce amyloid plaques in the brain, although it is still unclear whether this leads to a meaningful clinical benefit for people with the disease.

Eli Lilly's donanemab trial differed from Biogen's lecanemab one in that people stopped taking the drug once their amyloid levels had dropped below a certain threshold. "The rationale is, if the target is gone, why keep shooting?" Cummings says. According to the press release, about half of the trial participants were able to stop taking the drug in less than one year.



More Alzheimer's drugs head for FDA review: what scientists are watching

Diana Zuckerman, president of the National Center for Health Research, a non-profit think tank in Washington DC, worries that stopping the drug could cause the disease to rebound or worsen, as is the case with many psychiatric drugs. She warns that longer-term follow-up studies will be needed. "Any time you're doing anything that affects the brain, you really do have to be cautious," she says.

Eli Lilly also found that donanemab worked best in people whose brains contained only moderate levels of another protein, called tau, that is also associated with Alzheimer’s progression. The company had calculated its results among its 1,182 trial participants who had moderate tau levels, but said that the improvement was still statistically significant when they combined these patients with the 552 who had high levels of tau.

Brent Forester, a geriatric psychiatrist at McLean Hospital in Belmont, Massachusetts, says it’s “fascinating” that removing amyloid also affects tau: the relationship between the two proteins, and their respective roles in disease progression are not fully understood. “If we could understand that better, we might understand why removing amyloid might have a clinical effect,” he says.

## Bleeding and seizures

Like lecanemab, donanemab carries a high risk of side effects — particularly a set of conditions called amyloid-related imaging abnormalities (ARIA) that can lead to seizures and bleeding in the brain. Researchers think that by attacking amyloid plaques, the antibodies inadvertently weaken blood vessels in the brain, and the effects are especially pronounced among people who are taking anticoagulant drugs. Eli Lilly’s press release said that ARIA rates were several times higher in people who received donanemab than in those who received placebos, and three patients in the trial died after experiencing the condition.

“The side effect is the biggest concern for all of us right now,” says Forester, who led earlier trials of donanemab and is currently working on a lecanemab trial. He adds that people with mild cognitive impairment function fairly well, and that even three deaths might be enough to signal that the risk of side effects outweighs the benefit of taking the drug.



Could drugs prevent Alzheimer’s? These trials aim to find out

Questions also remain about information that is missing from the announcement, including whether donanemab worked at all among people who had high levels of tau. “This whole publication-by-press-release is really bad,” Zuckerman says.

Furthermore, the results that Eli Lilly released show only a slowing of cognitive decline relative to the placebo group, rather than how much donanemab affects the absolute rate of a person’s decline. It’s unclear, Zuckerman says, whether that difference is great enough to be noticeable to people with Alzheimer’s and their families.

With at least three monoclonal antibodies soon to be on the market, Mesulam worries that excitement around them will decrease drug companies’ enthusiasm for developing drugs for Alzheimer’s targets other than amyloid. “The next 20 to 25 years will be taken up by better amyloid drugs,” he says. The Alzheimer’s market is likely to be very lucrative for drug companies — lecanemab, for instance, costs more than US$26,000 per year of treatment — but Mesulam worries that the cost of Alzheimer’s drugs will strain the US health-care system.

Still, the initial results provide “further support that this therapy will have some role with the right patients at the right time in illness”, Forester says. “I’m cautiously optimistic.”

*Nature* **617**, 232-233 (2023)

*doi: https://doi.org/10.1038/d41586-023-01537-5*

# References

1. van Dyck, C. H. *et al. New Eng. J. Med.* **388**, 9–21 (2023).

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01537-5

| Section menu | Main menu |

- NEWS
- 04 May 2023

# Wuhan market samples fail to shed further light on COVID origins

New analysis of genomic data from market swabs highlights their limitations.

- Dyani Lewis




The spread of COVID-19 is thought to have begun at the Huanan market in Wuhan, China.Credit: Imaginechina Limited/Alamy

Samples collected at the Huanan Seafood Wholesale Market in Wuhan, China, in the early weeks of the COVID-19 pandemic are of limited value

for pinpointing which animal species — if any — infected people at the market, according to a new analysis[1].

Two previous analyses of the data described genetic material from various wild animals, suggesting it was possible that these animals could have passed the virus to people at the market. The new analysis attempts to identify the specific animal responsible for the spillover — but comes up empty.

“I would basically describe this as a negative result,” says Jesse Bloom, a virologist at the Fred Hutchinson Cancer Center in Seattle, Washington, who conducted the latest analysis, which has not yet been peer reviewed. “It doesn’t say either way whether there were ever any infected animals in the market,” he says. “It doesn’t say what the origin of the virus is.”

Florence Débarre, an evolutionary biologist at the French national research agency CNRS in Paris, who co-authored the first report on the animal sequences says that the latest analysis could never have answered the question of which animal hosts shed the viral material detected, because there was so much virus being shed by humans by the time the samples were collected.

Bloom analysed data on genetic material detected in swabs collected at the Huanan market by scientists at the Chinese Center for Disease Control and Prevention (China CDC) after the market was closed on 1 January 2020 owing to concerns about an emerging virus, which was later named SARS-CoV-2. In February 2022, scientists at the China CDC published a preprint paper[2] that revealed the presence of SARS-CoV-2 in a small number of environmental samples — from stalls, floors and walls — taken at the market. But the China CDC did not report on animal sequences detected in those samples and did not release the underlying data.

Bloom’s analysis is the third study in two months to focus on the genetic sequences of animals in the samples. The first report — posted in March by an international team of researchers including Débarre who discovered a subset of the China CDC data on GISAID, an online public data repository — found evidence of wild animals in samples positive for SARS-CoV-2[3]. The genomic sequences identified included those of raccoon dogs

(*Nyctereutes procyonoides*) and hoary bamboo rats (*Rhizomys pruinosus*) — species that could have been intermediate hosts of the virus. In the second, published in *Nature* in April, scientists at the China CDC provided their own analysis[4] of the animal sequences. The list of animals was different, but did include notable species such as bamboo rat and raccoon dog. They also posted the complete data set online.

Virologist George Gao at the China CDC, and lead author of the *Nature* study, says the new analysis yielded results similar to the China CDC's work.

## Signs of infection

Neither of the two previous reports on the animal sequences could confirm that animals were infected with SARS-CoV-2. Bloom's analysis also supports these results.

Bloom investigated whether environmental swabs rich in viral sequences were associated with genetic material from a particular animal, which could be a sign of infection. But there was no such association that made sense, says Bloom.

In fact, the strongest associations were with species, such as fish, cows and goats, that SARS-CoV-2 is not known to infect, says Bloom. "This doesn't provide any more insight on whether raccoon dogs, or any other animal there, were or were not ever infected with SARS-CoV-2," he says. The swab data merely confirm that the virus was widespread at the market, he says.

"We always knew that we could not definitively say if raccoon dogs were carrying [the virus] or if there had just been a commingling of the DNA," says Alice Hughes, a conservation biologist at the University of Hong Kong, who was not involved in any of the studies.

Making inferences based on precise amounts of genetic material found in environmental samples is problematic, says Hughes. "Some species are just going to be shedding more DNA than others," she says, because of the animals' size, their behaviour or how they are handled and processed at the

market. Sequencing can also introduce bias, says Hughes, because some genetic sequences amplify better than others.

Still, Débarre emphasizes that the latest analysis is important, because it independently confirms that specific wild animals capable of being infected with SARS-CoV-2 were at the market before its closure. It also shows that those animals were located in the part of the market linked to the most human cases, which, she says, adds weight to the hypothesis that the pandemic had a natural origin.

## More samples needed

Débarre says that more could be learnt if any of the original samples remain. For example, targeted sequencing of the virus might give a clearer picture of the early evolutionary history of the virus.

Hughes says that frozen animal carcasses or blood samples collected from people in late 2019 could hold some answers if any such samples exist. New surveys looking for SARS-CoV-2 antibodies in blood collected from people working in the supply chain of the wild-animal trade are unlikely to be useful, she says, because of the subsequent widespread SARS-CoV-2 infections in the Chinese population.

Bloom says that the existing swab data might yet yield further insights. For example, the sequences could reveal which wild or farmed populations the animals at the market came from. That information could be used to target geographical locations that researchers should investigate for the presence of coronaviruses related to SARS-CoV-2.

Gao says that the China CDC has no additional data from the early swabs. “We do not have anything of the earlier sequencing data from China CDC that have not [been] shared with the world,” he says. “The origin of the virus is a scientific question and remains elusive.”

*Nature* **617**, 233-234 (2023)

*doi: https://doi.org/10.1038/d41586-023-01483-2*

# References


1. Bloom, J. D. Preprint at bioRxiv https://doi.org/10.1101/2023.04.25.538336 (2023).

2. Gao, G. *et al.* Preprint at Research Square https://doi.org/10.21203/rs.3.rs-1370392/v1 (2023).

3. Crits-Christoph, A. *et al.* Preprint at Zenodo https://doi.org/10.5281/zenodo.7754299 (2023).

4. Liu, W. J. *et al. Nature* https://doi.org/10.1038/s41586-023-06043-2 (2023).


---


This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01483-2

- NEWS
- 03 May 2023
- Correction 04 May 2023

# ‘A good day’: FDA approves world’s first RSV vaccine

Green light from US regulatory agency for GSK’s jab caps off a decades-long quest.

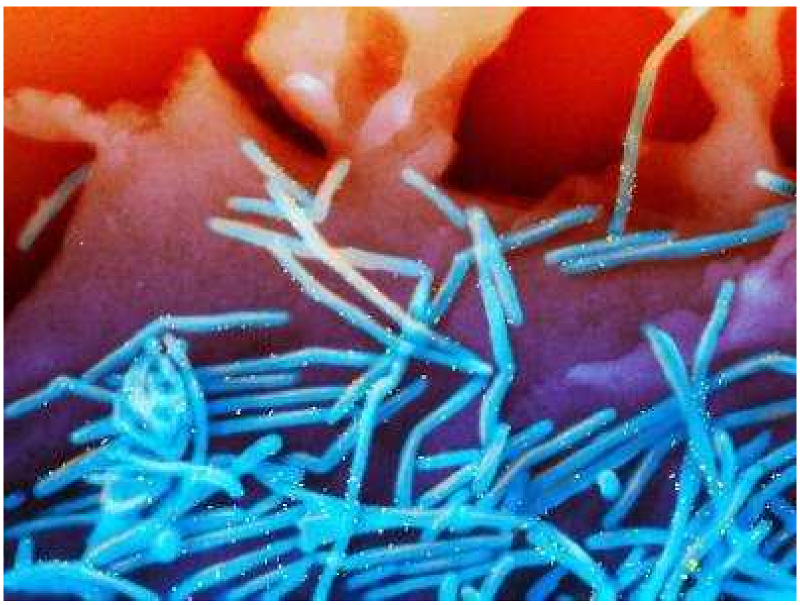
- Myriam Vidal Valero




The pharmaceutical company GSK beat competitors when it had its respiratory syncytial virus (RSV) vaccine approved first.Credit: iStock/Getty

The US Food and Drug Administration (FDA) has approved GSK's vaccine against respiratory syncytial virus (RSV) for use in people aged 60 years and older. This is the first RSV vaccine to gain approval anywhere in the world, and researchers are celebrating.

"It's a very big deal to have options available to prevent RSV disease," says Barney Graham, senior adviser for global-health-equity trials at Morehouse School of Medicine in Atlanta, Georgia.



Will a new wave of RSV vaccines stop the dangerous virus?

RSV usually causes mild symptoms similar to those that arise during a common cold. But for older people, it can be deadly. According to the US Centers for Disease Control and Prevention (CDC), each year RSV kills approximately 6,000 to 10,000 adults in the United States who are 65 years or older, and sends 60,000 to 160,000 to hospital. People in this age group with comorbidities such as chronic obstructive pulmonary disease, asthma or congestive heart failure are especially at risk.

The FDA approved GSK's RSV vaccine, to be sold as Arexvy, on the basis of phase III clinical-trial data submitted by the pharmaceutical company, which is based in London. Those data showed that the shot reduced the risk of people aged 60 and older developing lower respiratory tract disease from

RSV by 82.6% and that of developing severe disease by 94.1%. The results were published in February in *The New England Journal of Medicine*[1].

The approval is "a tremendous opportunity to help address a really important public-health need", says Leonard Friedland, vice-president and director of scientific affairs and public health for GSK's US vaccines division.

## A life's work

The FDA expedited the approval of GSK's vaccine last November, during a period when the United States was grappling with a 'tripledemic' of people with RSV, influenza or COVID-19 flooding hospitals.



The race to make vaccines for a dangerous respiratory virus

The technology underlying the RSV vaccine has been almost 60 years in the making. In the 1960s, during a clinical trial of an RSV vaccine, two of the participating children died and 80% ended up in hospital[2]. Understanding what had happened and finding a solution became Graham's life's work. "The first 20 years were spent primarily working out how to make a vaccine that could be safe," he says.

The original vaccine contained an inactivated RSV virus. Eventually, the vaccine-research community pivoted its strategy.

In 2008, Graham teamed up with Jason McLellan and Peter Kwong at the US National Institutes of Health in Bethesda, Maryland, and others to investigate the structural biology of RSV. Over time, they learnt that the virus used protein F, a molecule on its surface, to infect human cells. With this knowledge, protein-based RSV vaccines could be developed that would introduce the molecule into cells to produce an immune response.

But the team eventually found that early jabs of this type had been designed around the wrong form of protein F, a ‘postfusion’ version that arises after the virus and cell have already joined. The researchers discovered a way to instead target the correct form of protein F, which hasn’t yet fully fused with any cell and was found to elicit neutralizing antibodies. They published their findings in 2013[3].

Their work paved the way for companies such as GSK, Pfizer and Moderna to develop the RSV vaccines in the pipeline today. “This is my lifetime’s work. It’s very gratifying to see this finally happening,” Graham says. “It’s a good day for RSV.”

## The race continues

After the approval of GSK’s vaccine, scientists hope others will follow. “There are many patients all around the world who can benefit from vaccination,” Friedland says.



RSV wave hammers hospitals — but vaccines and treatments are coming

The pharmaceutical firm Pfizer, based in New York City, has also developed a protein-based RSV vaccine for people aged 60 and older that the FDA is expected to approve later this month. Biotechnology company Moderna in Cambridge, Massachusetts, has an mRNA-based vaccine for preventing RSV in adults in the same age group under expedited review at the agency.

Moreover, an FDA advisory panel will convene on 18 May to consider the safety and effectiveness of Pfizer's RSV vaccine for pregnant people. During a phase III clinical trial, pregnant people who received the jab gave birth to newborn babies who were then monitored for illness. The vaccine reduced the risk of infants up to 90 days old having a severe lower respiratory tract illness from RSV by 81.8%[4]. The FDA will decide whether to approve the vaccine by August.

Like older adults, babies are at high risk from RSV. "RSV seems like a harmless thing," but it's not, says Mina Suh, a scientist at the company EpidStrategies who is based in Irvine, California, that evaluates epidemiological studies. Each year, 58,000 to 80,000 children in the United States who are aged 5 years and under are hospitalized because of RSV and 100 to 300 die, according to the CDC.

Newborn babies' immune systems don't respond robustly to many vaccines, so it is challenging to give them shots directly. To protect infants, Pfizer opted to immunize pregnant people a few months before they gave birth. Their bodies make the antibodies that are transferred to their newborn.

But before this next tranche of vaccines is approved, GSK's jab will move ahead. It is expected to be approved in Europe soon. And the next step in the US process is for the CDC's Advisory Committee on Immunization Practices to weigh in. The panel will recommend who can receive the GSK jab, how it is given and when it will it be available to the public. The committee's next meeting is in June. This is an important step, Friedland says. "Vaccines don't save lives. It's vaccination that does."

*Nature* **617**, 234-235 (2023)

*doi: https://doi.org/10.1038/d41586-023-01529-5*

### Updates & Corrections

- **Correction 04 May 2023**: An earlier version of this story spelled researcher Jason McLellan's name incorrectly.

## References

1. Papi, A. *et al. N. Engl. J. Med.* **388**, 595–608 (2023).
2. Kim, H. Y. *et al. Am. J. Epidemiol.* **89**, 422–434 (1969).
3. McLellan, J. S. *et al. Science* **342**, 592–598 (2013).
4. Kampmann, B. *et al. N. Engl. J. Med.* **388**, 1451–1464 (2023).

---

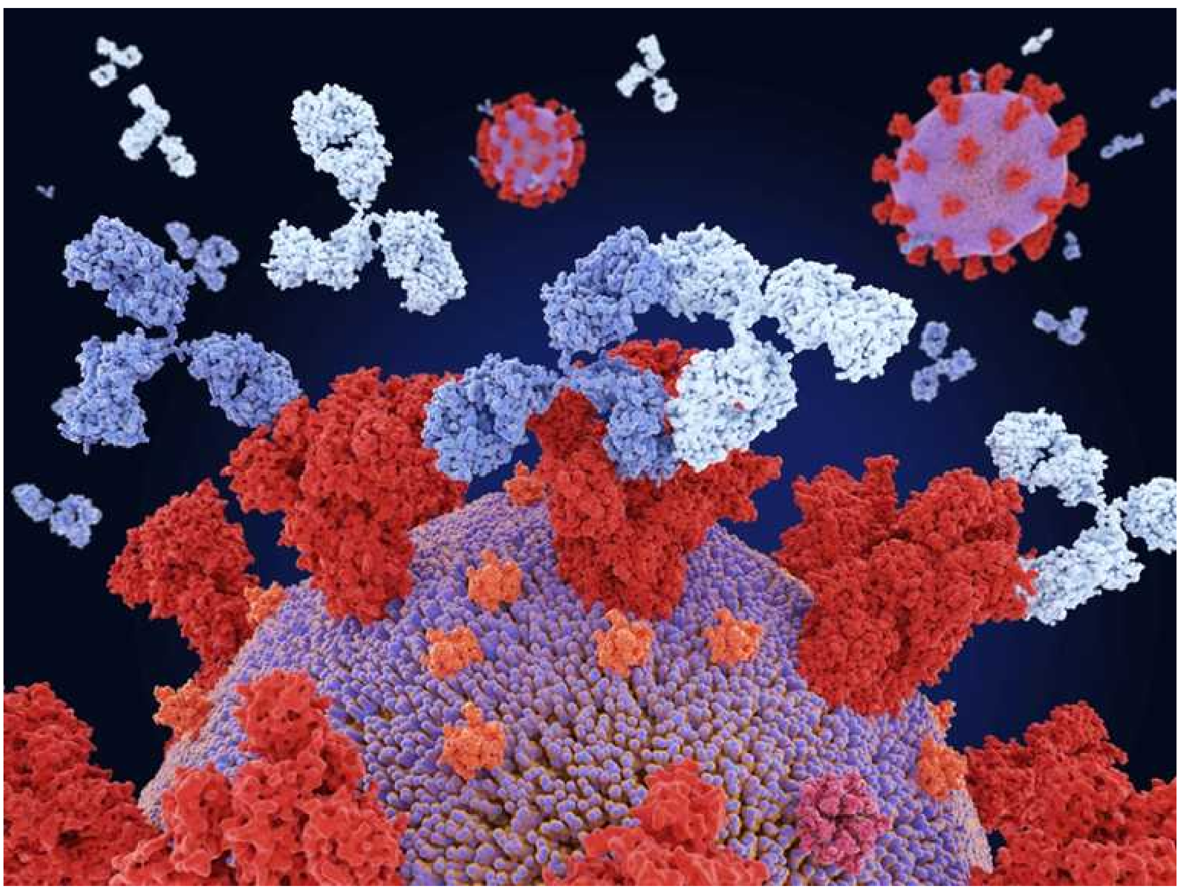
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01529-5

- NEWS
- 04 May 2023

# How generative AI is building better antibodies

Language models similar to those behind ChatGPT have been used to improve antibody therapies against COVID-19, Ebola and other viruses.

- Ewen Callaway



Monoclonal antibodies (y-shaped) binding to sites on a SARS-CoV-2 virus spike protein (red) (artist's conception).Credit: Juan Gaertner/Science Photo

Library

At the height of the pandemic, researchers raced to develop some of the first effective treatments against COVID-19: antibody molecules isolated from the blood of people who had recovered from the disease.

Now, scientists have shown that generative artificial intelligence (AI) can provide a shortcut through some of this laborious process, suggesting sequences that boost the potency of antibodies against viruses such as SARS-CoV-2 and ebolavirus. A study published last week in *Nature Biotechnology*[1] is part of growing efforts to apply 'neural networks' similar to those behind the ChatGPT AI platform to antibody design.

Antibody drugs for diseases including breast cancer and rheumatoid arthritis bring in more than US$100 billion in worldwide sales each year. Researchers hope that generative AI — neural networks that can create text, images and other content on the basis of learnt patterns — will speed up development and help to unlock antibody drugs for targets that have resisted conventional design approaches.

"There's intense interest in discovering and engineering antibodies, and how one makes antibodies better," says Peter Kim, a biochemist at Stanford University in California, who co-led the *Nature Biotechnology* paper.

## Immune weapons

Antibodies are among the immune system's key weapons against infection. The proteins have become a darling of the biotechnology industry, in part because they can be engineered to attach to almost any protein imaginable to manipulate its activity. But generating antibodies with useful properties and improving on these involves "a lot of brute-force screening", says Brian Hie, a computational biologist at Stanford who also co-led the study.

To see whether generative AI tools could cut out some of the grunt work, Hie, Kim and their colleagues used neural networks called protein language models. These are similar to the 'large language models' that form the basis

of tools such as ChatGPT. But instead of being fed vast volumes of text, protein language models are trained on tens of millions of protein sequences.

Other researchers have used such models to design completely new proteins, and to help predict the structure of proteins with high accuracy. Hie's team used a protein language model — developed by researchers at Meta AI, a part of tech giant Meta based in New York City — to suggest a small number of mutations for antibodies.

The model was trained on only a few thousand antibody sequences, out of the nearly 100 million protein sequences it learnt from. Despite this, a surprisingly high proportion of the model's suggestions boosted the ability of antibodies against SARS-CoV-2, ebolavirus and influenza to bind to their targets.

Alterations to a therapy approved to treat Ebola and a COVID-19 treatment bettered these molecules' ability to recognize and block the proteins these viruses use to infect cells. (The COVID-19 antibody is not effective against Omicron and its subvariants, and the AI-guided changes are unlikely to restore effectiveness, Hie says.)

Many of the suggested changes to antibodies occur outside the regions of the protein that interact with its target, which are usually the focus of engineering efforts, says Kim. "The model is reaching to information which is completely, or largely, non-obvious to even the experts in antibody engineering," he adds. "To me that's the 'holy cow, what's going on here?' moment."

## Completely new proteins

"This is a tool people will use to improve their antibodies," says Charlotte Deane, an immuno-informatics researcher at the University of Oxford, UK. "I think it's really cool." But she adds that many researchers hope that, instead of simply improving existing antibodies, generative AI will be able to create entirely new ones that will bind to a target of choice.

This ability could help researchers to develop drugs for molecular targets that have resisted other antibody-design approaches, says Surge Biswas, co-

founder of Nabla Bio, a company in Boston, Massachusetts, that is working on the challenge.

For example, AI could help to tackle G-protein-coupled receptors, a family of proteins sandwiched into cell membranes that are involved in neurologic disorders, heart disease and myriad other conditions. Generative AI could also help in the design of antibody drugs that are able to latch onto multiple targets, such as a tumour protein and an immune cell that can the kill that tumour, says Biswas.

Possu Huang, a bioengineer at Stanford, says that protein language models are powerful and very good at optimizing existing proteins, including antibodies. But models trained only on protein sequences might struggle to come up with truly new antibodies that recognize a specified protein.

Researchers say they are making progress. In March, scientists at Absci, a biotech firm in Vancouver, Washington, reported what they say is a first step towards making new antibodies with AI, in a preprint posted on the bioRxiv server[2]. Using a model incorporating protein sequences, as well as experimental data, they generated new designs for several important regions of an antibody drug used to treat breast cancer.

A key challenge in designing completely new antibodies is that their ability to recognize a particular target depends on floppy loops in the antibody structure. These interactions have proved difficult to model with AI, say researchers.

Last year, Huang's team developed[3] a generative AI tool that can create proteins able to bind strongly to a specified target — in one case, snake venoms — using such loops. The same approach could help to make new antibodies, Huang says, but it might require more data than are currently available about how antibodies interact with their targets.

"I don't think anybody's really figured this out," adds Biswas.

*Nature* **617**, 235 (2023)

*doi: https://doi.org/10.1038/d41586-023-01516-w*

# References

1. Hie, B. L. *et al. Nature Biotechnol*. https://doi.org/10.1038/s41587-023-01763-2 (2023).

2. Shanehsazzadeh, A. *et al.* Preprint at BioRxiv https://doi.org/10.1101/2023.01.08.523187 (2023).

3. Eguchi, R. R. *et al.* Preprint at BioRxiv https://doi.org/10.1101/2022.12.22.521698 (2022).

---

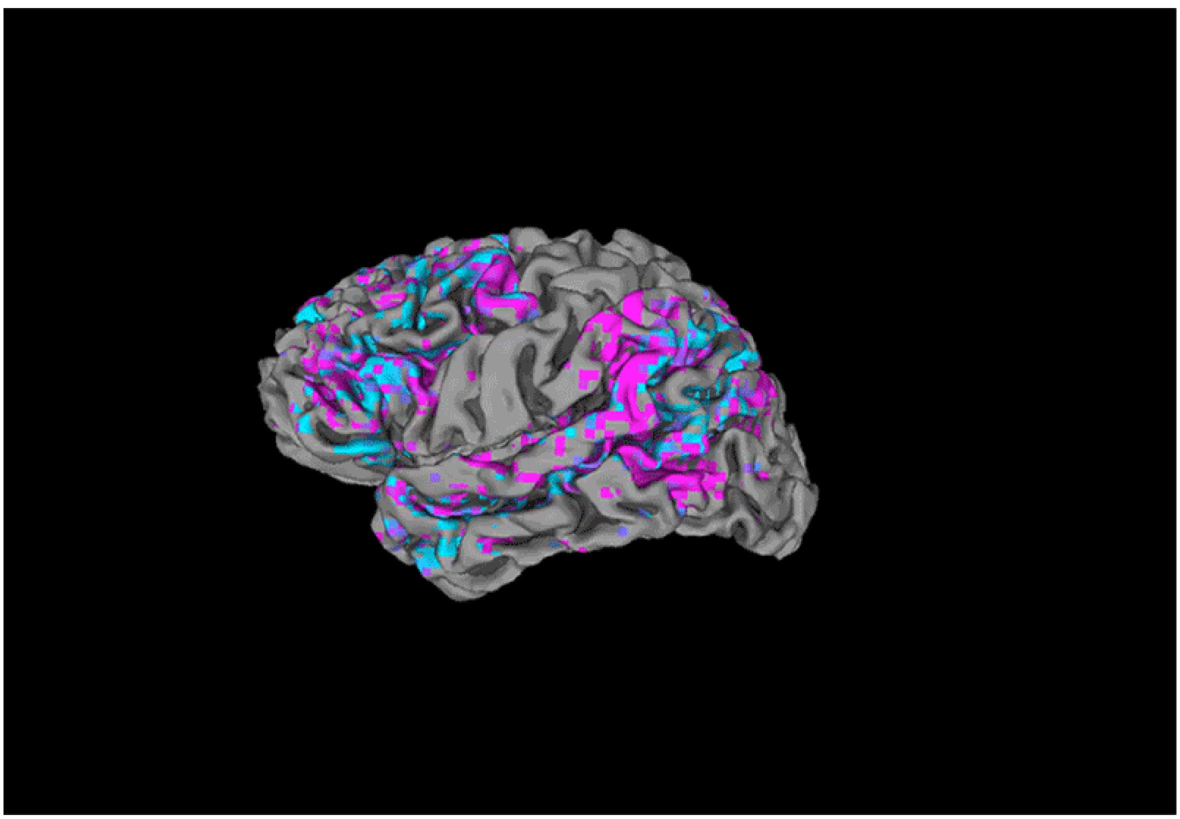
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01516-w

- NEWS EXPLAINER
- 02 May 2023

# Mind-reading machines are here: is it time to worry?

Neuroethicists are split on whether a study that uses brain scans and AI to decode imagined speech poses a threat to mental privacy.

- Sara Reardon



The study used functional magnetic resonance imaging (fMRI) to decode people’s thoughts.Credit: Jerry Tang and Alexander Huth

The little voice inside your head can now be decoded by a brain scanner — at least some of the time. Researchers have developed the first non-invasive method of determining the gist of imagined speech, presenting a possible communication outlet for people who cannot talk. But how close is the technology — which is currently only moderately accurate — to achieving true mind-reading? And how can policymakers ensure that such developments are not misused?

Most existing thought-to-speech technologies use brain implants that monitor activity in a person's motor cortex and predict the words that the lips are trying to form. To understand the actual meaning behind the thought, computer scientists Alexander Huth and Jerry Tang at the University of Texas at Austin and their colleagues combined functional magnetic resonance imaging (fMRI), a non-invasive means of measuring brain activity, with artificial intelligence (AI) algorithms called large language models (LLMs), which underlie tools such as ChatGPT and are trained to predict the next word in a piece of text.

In a study published in *Nature Neuroscience* on 1 May, the researchers got 3 volunteers to lie in an fMRI scanner and recorded the individuals' brain activity while they listened to 16 hours of podcasts each[1]. By measuring the blood flow through the volunteers' brains and integrating this information with details of the stories they were listening to and the LLM's ability to understand how words relate to one another, the researchers developed an encoded map of how each individual's brain responds to different words and phrases.

Next, the researchers recorded the participants' fMRI activity while they listened to a story, imagined telling a story or watched a film that contained no dialogue. Using a combination of the patterns they had previously encoded for each individual and algorithms that determine how a sentence is likely to be constructed on the basis of other words in it, the researchers attempted to decode this new brain activity. The video below shows the sentences produced from brain recordings taken while a study participant watched a clip from the animated film *Sintel*, about a girl caring for a baby dragon.

Credit: Jerry Tang and Alexander Huth

## Hit and miss

The decoder generated sentences that got the gist of what the person was thinking: the phrase ‘I don’t have my driver’s license yet’, for instance, was decoded as ‘she has not even started to learn to drive yet’. And it did a fairly accurate job of describing what people were seeing in the films. But many of the sentences it produced were inaccurate.

The researchers also found that it was easy to trick the technology. When participants thought of a different story while listening to a recorded story, the decoder could not determine the words they were hearing. The encoded map also differed between individuals, meaning that the researchers could not create one decoder that worked on everyone. Huth thinks that it will become even more difficult to develop a universal decoder as researchers create more detailed maps of individuals’ brains.

Determining how the brain creates meaning from language is enormously difficult, says Francisco Pereira, a neuroscientist at the US National Institute of Mental Health in Bethesda, Maryland. “It’s impressive to see someone pull it off.”‘

## ‘Wake-up call’

Neuroethicists are split on whether the latest advance represents a threat to mental privacy. “I’m not calling for panic, but the development of sophisticated, non-invasive technologies like this one seems to be closer on the horizon than we expected,” says bioethicist Gabriel Lázaro-Muñoz at Harvard Medical School in Boston. “I think it’s a big wake-up call for policymakers and the public.”

But Adina Roskies, a science philosopher at Dartmouth University in Hanover, New Hampshire, says that the technology is too difficult to use — and too inaccurate — to pose a threat at present. For starters, fMRI scanners are not portable, making it difficult to scan someone’s brain without their cooperation. She also doubts that it would be worth the time or cost to train a decoder for an individual for any purpose other than restoring communication abilities. “I just don’t think it’s time to start worrying,” she

says. “There are lots of other ways the government can tell what we’re thinking.”

Greta Tuckute, a cognitive neuroscientist at the Massachusetts Institute of Technology in Cambridge, finds it encouraging that the decoding system could not be applied across individuals and that people could easily trick it by thinking of other things. “It’s a nice demonstration of how much agency we actually have,” she says.

## Proceed with caution

Nevertheless, Roskies says that even if the decoder doesn’t work well, problems could arise if lawyers or courts use it without understanding its scientific limitations. For instance, in the current study, the phrase ‘I just jumped out [of the car]’ was decoded as ‘I had to push her out of the car’. “The differences are stark enough they could make an enormous difference in a legal case,” Roskies says. “I’m afraid they will have the ability to use this stuff when they shouldn’t.”

Tang agrees. “The polygraph is not accurate but has had negative consequences,” he said in a press conference. “Nobody’s brain should be decoded without their cooperation.” He and Huth called for policymakers to proactively address how mind-reading technologies can and cannot be legally used.

Lázaro-Muñoz says that policy action could mirror a 2008 US law that prevents insurers and employers from using people’s genetic information in discriminatory ways. He also worries about the implications of the decoder for people with conditions such as obsessive–compulsive disorder, who can experience unwanted, intrusive thoughts about harming people that they would never act on.

Pereira says the matter of how accurate decoders could become is an open question, as is whether they could eventually become universal, instead of being specific to an individual. “It depends on how unique you think humans are,” he says.

Although the decoder could eventually become good at predicting the next word in a series, it might struggle to interpret metaphors or sarcasm. There's a big step, Pereira says, between putting words together and determining how the brain encodes the relationships between the words.

*Nature* **617**, 236 (2023)

*doi: https://doi.org/10.1038/d41586-023-01486-z*

## References

1. Tang, J., LeBel, A., Jain, S. & Huth, A. G. *Nature Neurosci*. https://doi.org/10.1038/s41593-023-01304-9 (2023).

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01486-z

- NEWS FEATURE
- 10 May 2023

# ‘The best way to get it right is to listen to us’ — autistic people argue for a stronger voice in research

Despite broader acceptance of neurodiversity, autistic advocates and scientists are still fighting for a chance to set clinical and scientific priorities.

- Emiliano Rodríguez Mega



Yadira García Rojas joined a global coalition to help shape the research agenda on autism. Credit: Marian Carrasquero for *Nature*

In 2019, after finding out that her toddler, Luna, was autistic, Yadira García Rojas became fixated on the condition. She would spend her days reading websites and academic papers, trying to learn whatever she could to help her daughter, who doesn't speak and needs full-time care and support.

When the treatments started, they came as something of a shock. At a clinical centre for mental health in Mexico City, therapists would encourage Luna to talk by offering her strawberries, then take them away until she asked for them verbally. They would punish the girl when she flapped her hands or did not make eye contact, García Rojas remembers. And they would make her crawl through a play tunnel that she was terrified of — sometimes until she stopped crying.

"It was literal torture," says García Rojas. Most of the interventions, she says, seemed directed at making Luna and other children hide their autism — to be less themselves — no matter the cost in tears and screaming. "I always came out in crisis."

García Rojas stopped the therapy sessions shortly after COVID-19 hit Mexico in 2020. A year later, she received her own autism diagnosis. She started a local advocacy group aimed at reducing stigma against autistic people in Mexico and later joined an international coalition focused on shaping the research agenda, the Global Autistic Task Force on Autism Research (GATFAR). The group, made up of autistic scientists, physicians, therapists and advocates, aims to help change how academics see autistic people, which influences how physicians and therapists work with them, she says. "If there's no change there, there's no change anywhere."

Formed in 2022, GATFAR adds to a growing chorus of autistic and non-autistic advocates, including researchers, physicians and clinicians, who argue that autism research, writ large, is not serving the needs of autistic people and could, in fact, be causing harm.

"The research agenda is not set with autistic people's interest or well-being in mind," says GATFAR member Heini Natri, a computational scientist at the Translational Genomic Research Institute in Los Angeles, California.

She and others say that despite growing awareness of autism and acceptance of neurodiversity in many parts of society, the academic and clinical research community still generally considers autism a disease, a disorder, a problem to be solved. The focus of research and therapies, they say, is often on suppressing autistic traits and adopting neurotypical behaviours, rather than developing services and programmes to support autistic people.

At the end of 2021, an influential report in *The Lancet* setting clinical research priorities for the next five years angered many autistic advocates because of its focus on studying medications and behavioural therapies, and because it adopted the term "profound autism" to describe autistic people who require round-the-clock care, and who are often non-speaking, minimally verbal or intellectually disabled[1].



High-profile autism genetics project paused amid backlash

It reignited a debate over who gets to set the agenda. Pushing back against GATFAR and other activists were researchers and parents arguing that autistic people who are non-speaking or intellectually disabled — and their families — are at risk of being left behind and forgotten about. They've likened some of the arguments made by autistic advocates to censorship that could drive young scientists out of the field. "Researchers themselves need to push back against these attempts to muzzle them and thus control the direction of autism research," wrote Amy Lutz, vice-president of the US

National Council on Severe Autism in Philadelphia, Pennsylvania, in an opinion article in January (see go.nature.com/3vtycmn).

But such warnings are groundless, says Mary Doherty, a member of GATFAR who is based in Navan, Ireland, and founded Autistic Doctors International, which represents more than 700 autistic physicians from around the globe. Autistic people have historically had no control over what non-autistic researchers say and do, she says, "but it's a narrative that suits a certain part of the research community".

More autistic and non-autistic researchers have started working together in studies that prioritize the views of autistic people. For the past decade, some autistic people have been leading the research. And some prominent scientists have begun to adjust their approach.

"Neurotypicals and autistics, we're both important," says García Rojas. "If we work together, we'll be better off."

# Silent subjects

While working towards a master's degree in psychology in 2016, Monique Botha regularly thought about quitting. Every other paper they read seemed to contain demeaning descriptions of people like them. Language deficits in autistic people stem from a "failed domestication of the human brain", one read[2]. Another concluded that autistic children lack the uniquely human ability of sharing emotions, experiences and activities, much like non-human apes[3].

"I got exposed to that and I was like, 'You know what? I think I might be out,'" says Botha, who now works as a community psychologist at the University of Stirling, UK, and studies how autism research dehumanizes, objectifies and stigmatizes autistic people. (Elliot Murphy, a cognitive neuroscientist at the University of Texas Health Science Center in Houston and a co-author of the paper on failed domestication[2], says that the concept is technical and distinct from colloquial understandings of domestication. He adds that he has reached out to Botha and other authors "to correct them, and I have never received a response".)

Despite being discouraged, Botha says, one study, in particular, convinced them to continue their research. It showed that autistic adults die on average 16 years earlier than the general population, mostly from suicide[4]. For those with an intellectual disability, it was 30 years earlier.

“My community is bleeding, my community is dying,” says Botha. In 2018, they and their adviser explained why, in a study that included more than 100 autistic participants. Internalized stigma and everyday discrimination, they found, built up stress that resulted in poor mental health[5]. Botha says research is partly to blame. “If you continually describe autistic people through subhuman terms, then you will allow for subhuman treatment.”



These labs are remarkably diverse — here’s why they’re winning at science

Dehumanizing language has shaped autism research for decades. Some of the most disparaging remarks came from Ole Ivar Løvaas, a Norwegian American clinical psychologist. During a 1974 interview, Løvaas referred to autistic children — his patients — as “little monsters” who were “severely disturbed”. “You have a person in the physical sense — they have hair, a nose and a mouth,” he said, “but they are not people in the psychological sense.”

Løvaas, who died in 2010, founded applied behaviour analysis (ABA), a therapeutic approach that aimed to make autistic kids indistinguishable from

their peers. Løvaas was also a proponent of gay and transgender ‘conversion therapy’. Today, ABA therapies remain the most common type of autism intervention covered by insurance in the United States. The strategies used to modify children’s behaviour have mostly improved — few providers still use spanking or electric shocks. But even though studies indicate that the therapies can teach communication and socialization skills[6], the approach has been scrutinized for the lack of solid supporting data. Some research even points to the possibility that they can lead to trauma and abuse — including a link to developing post-traumatic stress symptoms[7].

“That’s what autistic people have been saying for decades,” explains Doherty. “But we’re not believed. We’re not seen as credible agents of knowledge about our own lives.”

It could be a legacy of the field. Conventionally, autistic people have been silent subjects in autism research, says Botha. “This has been thought of as acceptable and, worse, it’s been seen as desirable.” Data indicate that negative attitudes towards autism are very much alive in the research community.

Late last year, Botha co-authored a study that interviewed about 200 autism scientists, both senior and in training, to investigate how they understand the people they study. Nearly 60% of them openly expressed ableist views[8] that stigmatized, dehumanized and objectified autistic people. One participant described them as “shut down from the outside world”, “rigid”, “emotional” and, oddly, “fat”. Medical views that autism is a disorder or disease correlated with higher odds of ableist ideas, the researchers found. In turn, involving autistic people in research strongly correlated with lower odds of ableism.



Tobi Abubakare says that the research agenda can be hostile to autistic people.Credit: IU Bloomington

The conclusions don’t surprise Tobi Abubakare, an autistic clinical psychologist at Indiana University Bloomington. “It is a hostile environment,” they say.

Botha thinks that the findings should be explored more, but that they “really get to the heart of the issue, which is power, including the power involved in creating autism knowledge”.

## Here to help

About four years ago, Kristen Bottema-Beutel, an autism researcher at Boston College in Massachusetts, went through a personal crisis. She had heard some autistic researchers and advocates criticize her field as ableist, and she couldn’t shake the feeling that she hadn’t stopped to inspect her own

work. Were her descriptions of autism harmful? Had she herself been ableist?

She felt that a lot of her colleagues were resistant to such reflection. “There’s this tendency to be like, ‘Well, I’m beyond critique because I’m just here to help.’” With that, she says, comes a lack of scrutiny for many of the treatments and interventions that are offered to autistic people.

In 2020, for example, a meta-analysis explored 1,700 studies on non-pharmacological interventions — from sensory and behavioural therapies, similar to the ones García Rojas took Luna to, to some that are animal-assisted or technology-based. The results echo what several others have found: most autism interventions rely on low- to moderate-quality evidence[9].

“When you actually look at the evidence, it is not particularly strong,” says Bottema-Beutel.

Although she was not involved in that work, she participated in another meta-analysis that failed to find solid methodological rigour in therapies for young children[10]. She and her colleagues found that only about one-third of studies that test behavioural interventions are randomized controlled trials — the gold standard for determining how effective a treatment is.

In other studies[11–13], Bottema-Beutel has found that many papers fail to disclose important conflicts of interest or to report unintended negative side effects, such as children becoming distressed, hurt, bored or fearful as a result of an intervention.

The attitude is almost like, ‘nothing that could result from our intervention would be worse than being autistic’, Bottema-Beutel says. After all, “we’re just trying to help”.

## A rising tide of change

When Christina Nicolaidis, a health-services researcher at Portland State University in Oregon, started researching partner violence in the 1990s,

mentors told her that people who had experienced domestic violence “hated” researchers. Nicolaidis was sceptical that they really felt this way. “That tells me that researchers have screwed up,” she says.

After speaking to advocates and survivors, she realized that their views were rarely considered in the research. That seemed misguided: they had a wealth of knowledge and experience that “no matter what I could learn in my classes, I couldn’t get”, Nicolaidis says.

The experience shaped her career. Now she and her colleague, autistic researcher Dora Raymaker, co-direct the Academic Autism Spectrum Partnership in Research and Education (AASPIRE) — one of the longest-standing projects in which autistic and non-autistic scientists, as well as community members, work together as equals.

Since 2006, “we’ve kept that initial desire to say, ‘Hey, we want to do research that is important to the autistic community’,” says Nicolaidis.

It’s a different way of approaching a field in which non-autistic perspectives have typically driven the decisions. Community-based participatory efforts see autistic people as a marginalized group, she says, rather than as patients whose brains have gone awry.




Rachel Kripke-Ludwig helps to ensure that autism research is relevant to autistic people.

In the conventional approach, several researchers “are mostly working off the wrong set of assumptions”, writes Rachel Kripke-Ludwig, a non-speaking autistic advocate and student based in Menlo Park, California. “The best way to get it right is to listen to us.” She, along with other autistic people, sits on AASPIRE’s community council, discussing whether research questions are relevant, recruitment protocols respectful and methods scientifically sound.

Earlier this year, AASPIRE launched a project to understand what outcomes matter most to autistic people — which could go beyond standard life achievements, such as marriage or employment.

Studies show that autistic people want more research on mental health, services and the other conditions that many of them experience — including

epilepsy, gastrointestinal pain, intellectual disabilities and sleep problems[14].

However, most funding goes to studies on genetic and environmental risk factors, treatments and interventions. And according to Steven Kapp, an autistic developmental psychologist at the University of Portsmouth, UK, "there's little long-term [data] that clearly indicate that autism research, as a whole, has benefited autistic people". The field, he says, "has not served the needs of autistic people well".

Even participatory approaches might not be a silver bullet, however. Some efforts to be inclusive end up being superficial and tokenistic, warns developmental psychologist Liz Pellicano at University College London.

"That's kind of critical," Pellicano says. "To learn and actually take on people's different expertise, there needs to be that shift" of power.

## Profound divisions

Feeding controversy is the 2021 *Lancet* report. The authors, an international commission of mostly non-autistic researchers and clinicians, drafted clinical and research recommendations for what should be done to address the needs of autistic people in the next five years[1].

Critics agreed with some of the report's calls, such as tackling inequality and producing better data on the effectiveness of therapies. But the document, they said, misrepresented the views of some autistic people by prioritizing studies on interventions and medications — as opposed to looking into causes of death, mental-health support, assisted communication or debunking pseudoscientific treatments.

"If autistic people [had been] included in a meaningful way," Natri says, "that five-year plan that the *Lancet* commission proposed would probably look very different."



How three research groups are tearing down the ivory tower

Tony Charman, a developmental and clinical psychologist at King's College London and co-chair of the commission, agrees that the report fell short in not including more autistic perspectives. "When I look back and read what we wrote, quite a lot of the research recommendations are coming from the dominant — but not the only — position within the group," he says. "There is a piece missing there."

The commission also adopted the term 'profound autism'. This drew pushback from autistic researchers and advocates[15], and from organizations, such as GATFAR[16], who saw the label as arbitrary and scientifically inaccurate. They argued that it unnecessarily splits the autistic community, undoing a decision in 2013 to merge a wide range of conditions, including Asperger's syndrome, into one diagnosis.

"It's interesting that it has created such a furore," says clinical psychologist Catherine Lord at the University of California, Los Angeles, who co-chaired the *Lancet* commission, adding that the report intended to call more attention "to people who are often forgotten".

Indeed, the proportion of autism studies with participants who would fit the commission's definition of 'profoundly autistic' has dramatically decreased over time[17] — partly, some argue, because of the fusing of autism diagnoses.

Alison Singer, a co-author of the *Lancet* report and president of the Autism Science Foundation, a New York City-based organization that funds autism studies, says that the term is necessary. “You have people with autism who are so high-functioning that they graduate from Harvard Law School. And then you have people like my daughter”, who has severe intellectual disability. “For me, the solution is we have to return to bifurcating the diagnosis,” she says.

Many autistic people see that as a step back to labels that they have rejected. “I am profoundly gifted, not profoundly low-functioning,” writes Payam, a non-speaking autistic advocate who is based in Atlanta, Georgia. Payam is not an exception, says his mother, Parisa Khosravi. “We need to presume competence and listen to our non-speakers,” she explains, “rather than assume intellectual disability.”

Many other autistic people who are non-speaking or have intellectual disabilities have found ways to speak up for themselves, says Zoe Gross, director of advocacy for the Autistic Self Advocacy Network in Washington DC. “It is completely inaccurate to say that as a group, autistic people with intellectual disabilities, or nonspeaking autistic people, can’t advocate for themselves,” she wrote in an e-mail. “Not all autistic people have access to a communication method that works for them, and for some people, the currently available communication methods may just not work.”

A further disagreement has fractured the autism community. Late last year, Singer co-authored a commentary saying that calls for autism researchers to be more careful about the language they use was ‘censorship’[18]. The piece cautioned that the concerns raised by self-advocates will push early-career investigators out of the field. “They may move on to study other disorders,” says Singer. “There is a real threat to research.”

The article generated sharp criticism from both autistic and non-autistic researchers. “That’s a wild claim,” says Bottema-Beutel. “It expresses a level of fragility that I think is inappropriate for a community of researchers.”

Botha agrees. “It’s laughable to pretend that autistic people have institutional and systemic power to steer autism research when traditionally we’ve been

kept out of it."

It's hard to say how influential the *Lancet* report, or the recommendation to use 'profound autism', will be, several sources told *Nature*. But the intense divisions the report has generated mask some crucial progress over the past few years.

For some, the debate has served as a wake-up call. "It's a kind of revolution, I think," says psychologist Simon Baron-Cohen at the University of Cambridge, UK.

In September 2021, he and his colleagues paused the Spectrum 10K study, a massive project aiming to understand the genetics of autism and co-occurring conditions. Critics argued that the study had commenced without meaningful consultation of autistic people and that the data it collects could be misused. The team has now launched a thorough consultation, to be finished later this month, that will probably alter the project's nature and protocols, Baron-Cohen says. "We're in listening mode at this point."

Autism research funders in Australia have, for some time now, been allocating more grants that reflect autistic priorities[19]. Other institutions, such as the Interagency Autism Coordinating Committee, a US federal advisory panel, have followed suit: its 2021–22 strategic plan for autism research called for more funding on well-being and inclusion.

In 2020, the International Society for Autism Research created a committee of autistic researchers — including Abubakare, Kapp and Raymaker — to foster collaborations with their non-autistic peers and offer mentorships to young autistic trainees starting to navigate autism research.

All these efforts fuel the hope that momentum to include autistic perspectives in autism science will continue to grow.

The only way forward, says Kripke-Ludwig, is to make room for more autistic people in academia — even if discomfort ensues. "I am not a broken version of normal," she wrote to *Nature*. "I am autistic, disabled and fabulous. Get used to it."

*Nature* **617**, 238-241 (2023)

*doi: https://doi.org/10.1038/d41586-023-01549-1*

## References

1. Lord, C. *et al. Lancet* **399**, 271–334 (2022).
2. Benítez-Burraco, A., Lattanzi, W. & Murphy E. *Front. Neurosci.* **10**, 373 (2016).
3. Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. *Behav. Brain Sci.* **28**, 675–691 (2005).
4. Hirvikoski, T. *et al. Br. J. Psychiatry* **208**, 232–238 (2016).
5. Botha, M. & Frost, D. M. *Soc. Ment. Health.* **10**, 20–34 (2020).
6. Leaf, J. B. *et al. J. Autism Dev. Disor.* **52**, 2838–2853 (2022).
7. Kupferstein, H. *Adv. Autism* **4**, 19–29 (2018).
8. Botha, M. & Cage, E. *Front. Psychol.* **13**, 105089 (2022).
9. Whitehouse, A. *et al. Interventions for Children on the Autism Spectrum: A synthesis of research evidence* (AutismCRC, 2020).
10. Sandbank, M. *et al. Psychol. Bull.* **146**, 1–29 (2020).
11. Bottema-Beutel, K. *et al. J. Child Psychol. Psychiatry* **62**, 5–15 (2021).
12. Bottema-Beutel, K. & Crowley, S. *Front. Psychol.* **12**, 676303 (2021).
13. Bottema-Beutel, K., Crowley, S., Sandbank, M. & Woynaroski, T. G. *Autism* **25**, 322–335 (2021).
14. Roche, L., Adams, D. & Clark, M. *Autism* **25**, 336–348 (2021).
15. Kapp, S. K. *Educ. Sci.* **13**, 106 (2023).

16. Pukki, H. *et al.* *Autism Adulthood* **4**, 93–101 (2022).

17. Stedman, A., Taylor, B., Erard, M., Peura, C. & Siegel, M. *J. Autism Dev. Disord.* **49**, 1378–1390 (2019).

18. Singer, A., Lutz, A., Escher, J. & Halladay, A. *Autism Res.* **16**, 497–501 (2022).

19. den Houting, J. & Pellicano, E. *J. Autism Dev. Disord.* **49**, 4400–4408 (2019).

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01549-1

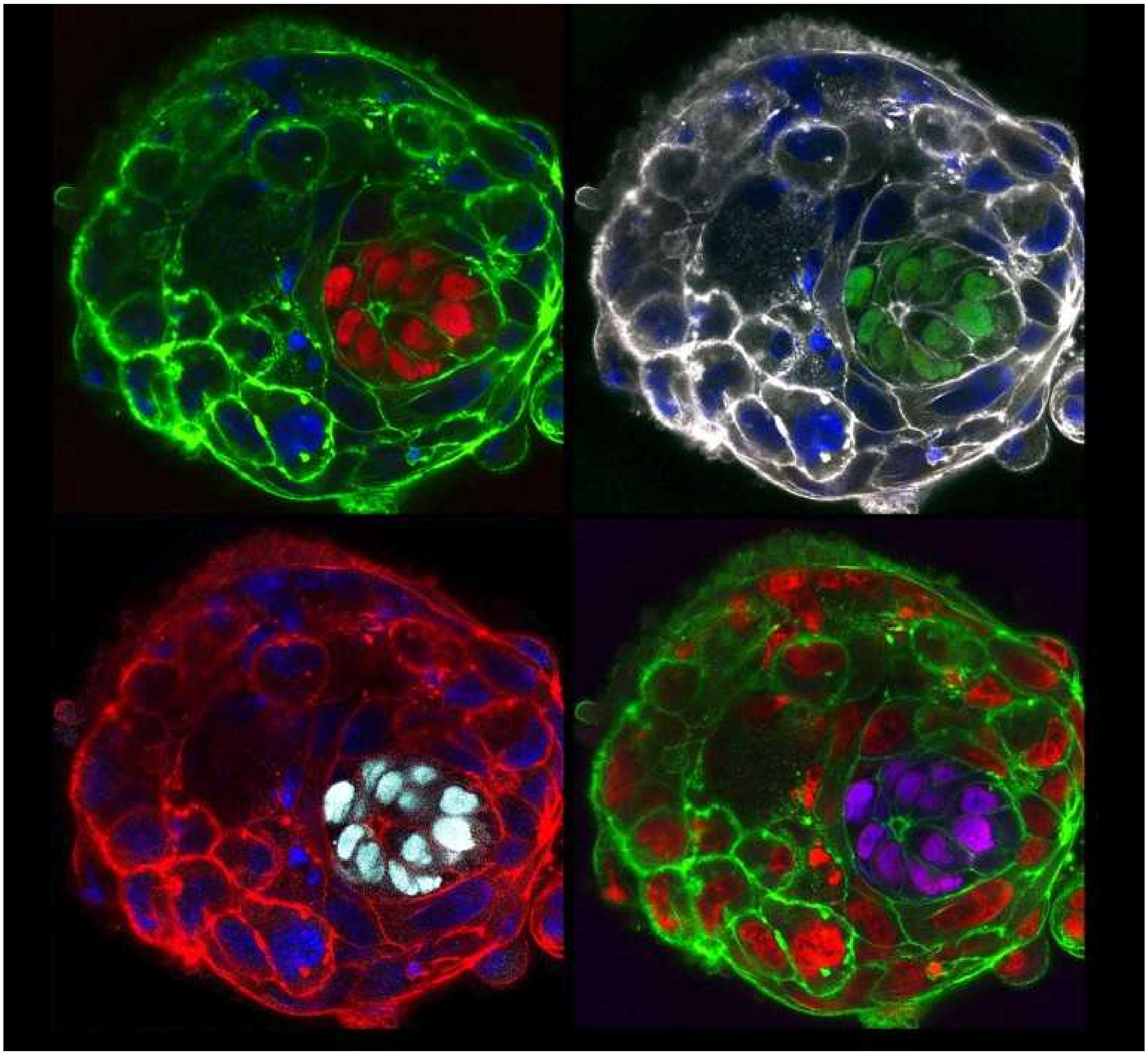
| Section menu | Main menu |

- NEWS FEATURE
- 10 May 2023

# Human embryo science: can the world’s regulators keep pace?

Concerned by the speed of research, policymakers are looking for alternative ways to establish rules on genome editing and other technologies.

- Philip Ball



Research on human embryos is governed by a patchwork of guidelines and rules.Credit: Zernicka-Goetz Lab/Cambridge University/SPL

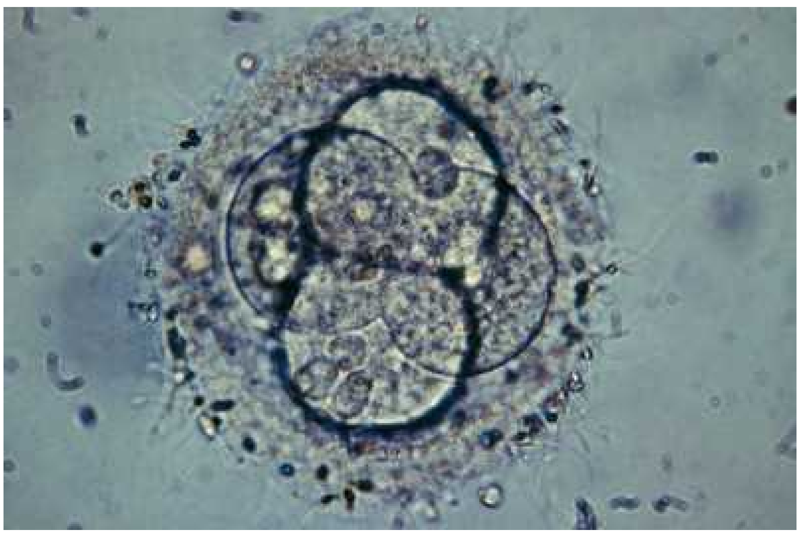
Outside the Third International Summit on Human Genome Editing, protesters held banners that read “Stop designer babies” and “Never again to eugenics”, calling for a ban on human genome editing. It was a reminder of the ethical quandaries that embryo science and genome editing can present.

Inside the meeting, held at the Francis Crick Institute in London in March, it was clear that the pace of the science prompting such concerns is quickening. Scientists reported, for example, that they had grown mice using eggs engineered from cells from male mice[1]. The following month, researchers in China reported[2] pregnancies in monkeys after transferring

artificial embryos, made from stem cells, into monkey wombs, although the pregnancies did not progress beyond a very early stage.

The conference also hosted a talk by a woman whose sickle-cell disease had been successfully treated by editing the genetic code in her body's blood-producing cells.

Regulators — whose job it is to establish the bounds of what is permissible — are struggling to keep up. "Regulatory approval cannot happen at the same speed as scientific development," says Peter Thompson, chief executive of the UK Human Fertilisation and Embryology Authority (HFEA), a government body seen as a trailblazer in regulating this branch of science and medicine.



Beyond CRISPR babies: how human genome editing is moving on after scandal

The problem was starkly demonstrated five years ago, when Chinese researcher He Jiankui announced, at the Second International Summit on Human Genome Editing in Hong Kong, that he had altered the genomes of twin girls born by *in vitro* fertilization (IVF), with the aim of giving them resistance to infection by HIV. He's work was widely condemned by researchers as irresponsible, and after a Chinese court judged it to involve 'illegal medical practice', he was given a three-year prison sentence.

The World Health Organization currently recommends against the use of heritable (or germline) human genome editing, but has no powers of enforcement — and neither does any other international body. Some national bodies, such as the HFEA, can create rules to govern research and its outputs, but such centralized oversight is relatively rare. Most scientists comply with guidelines formulated by professional bodies such as the International Society for Stem Cell Research (ISSCR), but these principles have neither legal force nor a mandate from wider society.

The ideal rulebook would be agile and able to evolve with the science, and it would include public and expert opinions. But it's not clear that existing procedures can do that job, so the HFEA is considering changing how it works to accommodate the pace of research. In February, it launched a public consultation on ideas to "future-proof" the Human Fertilisation and Embryology Act — the law that governs UK embryo and fertility research.

But will such changes be enough to keep up?

## Delicate decisions

Rules and laws governing embryo science have often emerged slowly and piecemeal. After the first IVF births in the 1970s, the UK government convened an expert panel to consider how embryo research should be regulated — for example, how long human embryos could be cultured and studied in the laboratory for. In 1984, the panel recommended a limit of 14 days from fertilization, a restriction that finally became law in the United Kingdom in 1990 and was adopted by several other countries. The ISSCR also adopted this limit in its guidelines — until 2021, when it proposed relaxing the restriction in some circumstances, in response to researchers' newfound ability to grow embryos right up to the limit and taking into account the potentially useful knowledge that could be gleaned from going beyond it.



Peter Thompson is head of the UK Human Fertilisation and Embryology Authority.Credit: HFEA

At the same time, the ISSCR advised against using methods such as CRISPR, a technique that cuts the DNA sequence at specific points, or potentially more accurate ‘base’ and ‘prime’ editing techniques, to alter the genome in human embryos, until the safety issues are better understood.

At the London summit, most scientists seemed to agree that somatic-cell editing, which creates changes that are not heritable, is warranted to treat conditions such as sickle-cell disease, provided that safety and efficacy are demonstrated.

The consensus is that germline editing requires much more cautious handling, because the changes are passed on to future generations. But if some of the new precision-editing methods are shown to be safe in human embryos, says Crick developmental and stem-cell biologist Robin Lovell-

Badge, one of the organizers of the summit, the whole field "could accelerate very fast" and demand for germline editing could increase.

## Nimbler rulemaking

Science moves fast — but regulation often cannot. For the HFEA, altering regulations would typically require a change in law. "The issue is the inevitable mismatch between a fixed legal framework, which is slow to change, and the constant development of the science," says Thompson.

One way to get around this, Thompson says, is to design 'primary' legislation such that it can be more easily amended by 'secondary' legislation, which does not require the same degree of parliamentary scrutiny or approval. That approach was successfully used by the HFEA to grant approval for a technique known as mitochondrial replacement therapy in 2016.



The mice with two dads: scientists create eggs from male cells

Another approach being explored by the HFEA is a 'regulatory sandbox': a kind of controlled experiment in which fresh ways of doing things are trialled in a small-scale real-world environment. This process was first developed in the finance sector for testing new and unproven investment tools. The HFEA's latest public consultation is a first step towards gauging people's appetite for such approaches.

Some scientific innovations might still fall between the cracks. Take embryo models: synthetic constructs, generally made from embryonic stem cells or stem cells derived from adult somatic cells, known as induced pluripotent stem cells (iPS cells), that recapitulate embryo development. There is no consensus on whether they should be regarded as genuine embryonic organisms or as types of tissue culture. Already, researchers have grown such embryo models *in vitro* until organs begin to develop[3], and the latest study on monkeys indicates that scientists can implant them and initiate pregnancies. Yet this is happening without clear regulations about what is permissible: they "are not regulated at all", says Lovell-Badge.

The HFEA has clear legal powers over some of these technologies, but in other countries regulation evolves more organically. In the United States, the government cannot regulate purely on moral grounds, but it has effectively barred research that creates or destroys human embryos, by restricting federal funding for it. Germline human genome editing is also blocked by an indirect bureaucratic gambit: clinical trials cannot happen without approval from the US Food and Drug Administration (FDA), but the FDA is prohibited from even considering such requests.

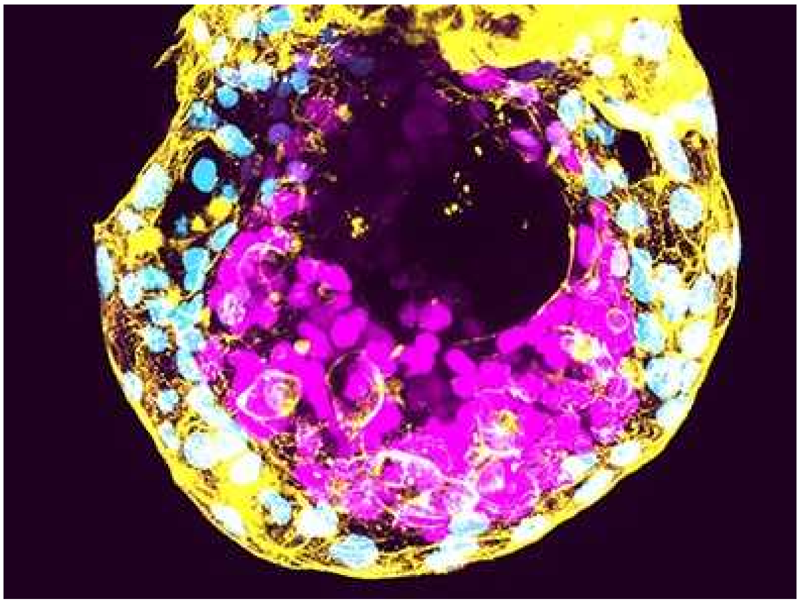
Meanwhile, US states can pass their own legislation controlling research and medical procedures, creating a patchwork. But bioethicist Alta Charo at the University of Wisconsin–Madison thinks there are advantages to such an ad hoc system. "We have a 'laboratory of the states'," she says. "It's very messy and inefficient, but it allows for exploration of options."

In Japan, bioethicist Tetsuya Ishii at Hokkaido University in Japan says that such technologies are constrained mostly just by guidelines that are "effective to some extent" but not legally enforceable, particularly for therapies using adult stem cells. By contrast, research and therapies based on iPS cells or embryonic stem cells get "a lot of oversight and regulation in Japan", says developmental biologist Cantas Alev at Kyoto University.

# Public opinion

When science moves fast, policymakers might respond reactively, says Bartha Knoppers, a specialist in the ethical and legal aspects of genomics

and biotechnologies at McGill University in Montreal, Canada. This mindset could unduly constrain science that has possible health benefits, she says. Asking those who might stand to benefit should be a key part of the decision-making. But "whether patients think regulation is going too fast or not fast enough depends on what condition you have", says Bettina Ryll, a founder of the Melanoma Patient Network Europe, who spoke at the genome-editing summit. People with a serious, progressive disease will be more willing to accept risks, she says.



[What's next for lab-grown human embryos?]

Consultations with stakeholders and the wider public can be difficult, costly — and slow. But without this input, there is the danger of mismatch between policy and public opinion. For instance, some existing restrictions, such as the US ban on federal funding for human-embryo research, have not been assessed in consultation with the public. "Does this [situation] represent public opinion?" says Charo. "No one knows."

Some researchers, however, question why time should be of the essence anyway. Sheila Jasanoff at Harvard University in Cambridge, Massachusetts, a specialist in the social and policy impacts of science and technology, pushes back against the idea that regulations must be changed to keep pace with a fast-moving field. "Why is the slow, deliberative process always seen as the problematic one?" she asks. With this philosophy in mind, in 2020,

she and bioethicist Benjamin Hurlbut at Arizona State University in Tempe launched the Global Observatory for Genome Editing, which aims to foster dialogue on this and related biotechnological developments between diverse groups of people.

Whatever the position about regulation, it pays to know what developments are coming. Some international and national organizations already conduct such ‘horizon scanning’ for fast-moving technologies — although it is not clear that any has yet considered embryo science. For example, the Organisation for Economic Co-operation and Development (OECD) runs an Observatory of Public Sector Innovation, which seeks to develop guidance for governments and the public sector on the opportunities and risks of emerging technologies such as artificial intelligence (another area in which advances are outpacing regulatory oversight).

Hank Greely at Stanford University in California, who researches the legalities of bioethics, has called for the wider establishment of such horizon-scanning groups, as well as auditing bodies that would keep tabs on whether regulations are being followed and remain fit for purpose[4]. The question is who would fund and oversee such projects, he adds; industries, governments and scientific institutions might all be considered to have vested interests.

In the end, making sure that fast-moving science is regulated in a timely and effective manner is about maintaining a robust social contract between science and society. “Are we continually asking ourselves, is this where we want to go? Who will it serve?” says Knoppers. “It’s about keeping vigil.”

*Nature* **617**, 242-243 (2023)

*doi: https://doi.org/10.1038/d41586-023-01550-8*

# References

1. Murakami, K. *et al. Nature* **615**, 900–906 (2023).

2. Li, J. *et al. Cell Stem Cell* **30**, 362–377 (2023).

3. Amadei, G. *et al. Nature* **610**, 143–153 (2022).

4. Greely, H. T. *Glob. Public Pol. Gov.* **2**, 266–282 (2022).

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01550-8

# Books & Arts

- **Ethics in outer space: can we make interplanetary exploration just?** [08 May 2023]
  • The prospect of settling the Moon, Mars and elsewhere requires urgent conversations about issues such as labour and reproductive rights far from Earth.

- BOOK REVIEW
- 08 May 2023

# Ethics in outer space: can we make interplanetary exploration just?

The prospect of settling the Moon, Mars and elsewhere requires urgent conversations about issues such as labour and reproductive rights far from Earth.

- Alexandra Witze



Future space explorers need to take care to avoid the harms caused by colonialism and exploitation on Earth.Credit: Detlev Van Ravenswaay/SPL

**Off-Earth: Ethical Questions and Quandaries for Living in Outer Space** *Erika Nesvold* MIT Press (2023)

**Reclaiming Space: Progressive and Multicultural Visions of Space Exploration** *James S. J. Schwartz, Linda Billings and Erika Nesvold (eds)* Oxford Univ. Press (2023)

From *Star Trek* to Apollo 17, space exploration is often framed as humanity pushing collectively towards a better future. But those utopian visions probably won't mesh with reality. The book *Off-Earth* explores the ethical implications of humans moving into outer space — and whether those who do can avoid bringing along Earthly problems such as environmental destruction and social injustice. *Nature* spoke to its author, Erika Nesvold.

Nesvold is a computational astrophysicist, game developer and a member of the team behind Universe Sandbox, a physics-based space simulator. Based in Severn, Maryland, she is also co-founder of the JustSpace Alliance, a non-profit organization that works for a more inclusive and ethical future in space, and co-editor of *Reclaiming Space*, a collection of essays that explores similar themes.

## Why is now a good time to talk about ethics in space exploration?

A lot of people are talking about these topics because of the growth of the private space-flight industry. For decades, human space exploration was done by national agencies with different motivations, different rhetoric surrounding it and different levels of public participation. With private space flight, members of the public — if they get rich enough — can actually think about going into space.



Erika Nesvold explores ethics of space travel.Credit: Todd Dring Photography

## You talk about ‘settling’ space, rather than colonizing it. Why?

Because of all the terrible behaviour that came out of the colonization model here on Earth. People talk about space as the final frontier — there’s always references to the Wild West. But living on the frontier in the Wild West was not an ideal experience for most, including Indigenous people, women and

Black people. We can't just pick and choose the nice shiny parts of history and apply them to space. We have to also look at what was harmful about those times and how we can avoid that in the future.

## What messy Earth problems could humans carry into space?

It's not just that we'll bring all the same problems with us into space. There are certain physical characteristics of the space environment that could make all of this worse — making people more vulnerable to exploitation, for example. People working in space — say, mining asteroids or whatever we send them out to do — are going to be in isolated environments without a lot of oversight or monitoring. They're going to be at the mercy of their employers for air, water, food and a trip back to Earth.

This has parallels with a lot of environments where we see labour exploitation and abuses on Earth. During my research for *Off-Earth*, I spoke to a labour-rights activist who talked about the Thai fishing industry, in which migrant workers are hired, their passports are taken away, they're put on boats and taken off to sea. And they can be kept out there for years and abused without anyone watching. They don't have a way back. Solutions that work on Earth, such as strong labour unions and regulation to defend workers' rights, will also help to safeguard future space workers — assuming we put deliberate effort into putting these protections into place.

## Why does *Off-Earth* also talk about reproductive rights in space?

This felt like one of the most futuristic topics in the book because nobody's saying, oh, let's start having children in space. But if we want someday to have a permanent human presence in space, then we have to be able to replenish our population without continually shipping people from Earth. That means human reproduction in space, but there are so many ethical complications.

In the long term, it leads into questions of population control. If you're in an environment of real scarcity, such as we imagine a space settlement would be, you need to keep your population from getting too big or too small. And we've seen on Earth that trying to control a country's population can lead to some really unethical and horrific practices.



Private companies are flocking to the Moon — what does that mean for science?

## What can people who have worked on these issues on Earth teach us?

Historians, especially colonial historians, can point to lessons learnt — cautionary tales and also success stories from the past. Anthropologists, sociologists and economists can explain how humans live together and how their culture shapes behaviour, and vice versa. And how all of this affects the success and happiness of communities.

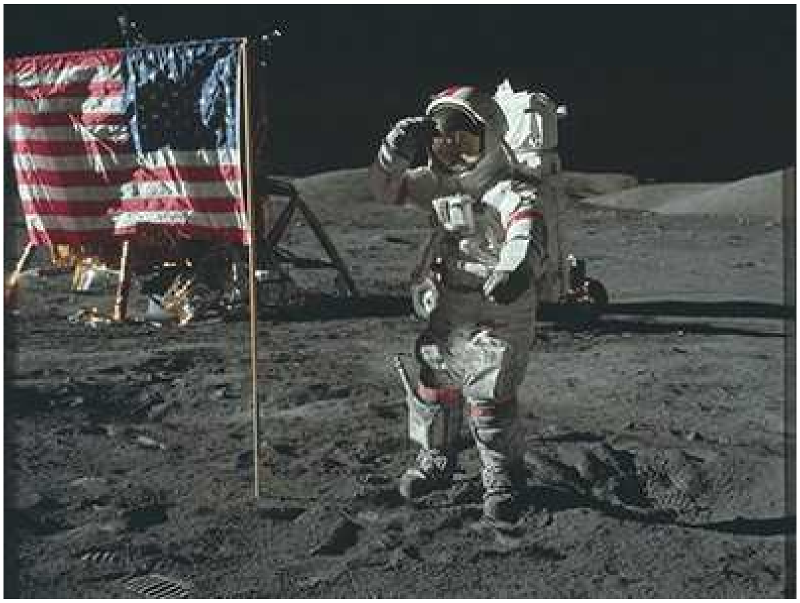
While researching my book I talked to Michelle Brown, a criminologist at the University of Tennessee, Knoxville, and Walidah Imarisha, who is a writer, an activist and director of the Center for Black Studies at Portland State University in Oregon. They both taught me a lot about prison abolition and restorative justice. If we talk about maybe not needing prisons in space,

then the question is, what’s the alternative? And it turns out there’s a ton of people who are exploring how different cultures around the world handle harm within their communities in ways that don’t involve locking people up in prison.

## How do we ensure that voices from all around the globe are heard?

There’s been discussion that we just need to bring more people to the table, to have more of these conversations across a diverse range of cultural backgrounds, expertise and lived experiences. But even that assumes it’s the Western space industry’s table, and that we get to decide who comes to it. What I’d rather see is more amplification of the conversations that people outside of the Western world view are having about space and their own societies — so that we can encourage a more global conversation that’s not dominated by one country or one cultural viewpoint.

A lot of people in the space industry like to talk about space as if they’re speaking on behalf of all of humanity, which is really disingenuous if you haven’t even consulted with all of humanity about their interests and what their motivations are for space. Someone living in a community whose culture is deeply tied to their land might be much more interested in how space technology could help them to continue protecting and restoring the health of their land and community, rather than imagining relocating to space in the distant future.



[Fifty years after astronauts left the Moon, they are going back. Why?]

## What impact does diversity among astronauts have on these issues?

Representation and diversity matters in space and in space travellers. NASA has committed to sending the first woman and the first person of colour to the Moon. It has the responsibility as a US government agency to reflect the demographics of its population and make sure the astronaut corps is as accessible for as many people as possible. In the long term, if you're looking at creating permanent populations living in space, you want them to be representative of humanity.

## What can researchers do to ensure the ethical exploration of space?

One thing I encourage astronomers to do is to learn from scientists in other disciplines, such as genetics, who have had to think about the ethical implications of their research and how they balance, on a personal level, the work and the potential harm it could cause. Having conversations involving different disciplines would be useful.

More broadly, scientists who want to help to build a better future in space but don't work in policymaking or philosophy can do a couple of things. They can have conversations with their friends and colleagues about what a better future in space would look like, what kind of world in space they would want to live in, and also just focus on making a better society here on Earth today, whatever corner of the globe they live in. If we do manage to make Earth a better place and a nicer society to live in, within our lifetimes, then we're helping our future in space.

*Nature* **617**, 245-246 (2023)

*doi: https://doi.org/10.1038/d41586-023-01551-7*

The author declares no competing interests.

This interview has been edited for length and clarity.

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01551-7

# Opinion

- **Humans and algorithms work together — so study them together** [10 May 2023]
  • Adaptive algorithms have been linked to terrorist attacks and beneficial social movements. Governing them requires new science on collective human–algorithm behaviour.

- COMMENT
- 10 May 2023

# Humans and algorithms work together — so study them together

Adaptive algorithms have been linked to terrorist attacks and beneficial social movements. Governing them requires new science on collective human–algorithm behaviour.

- J. Nathan Matias




YouTube and Twitter both use adaptive algorithms that recommend videos or suggest accounts to follow. Credit: NurPhoto via Getty

Later this year, the US Supreme Court will rule on a case with profound implications for society and technology. The case centres on Nohemi Gonzalez, a US student who in 2015 was shot and killed by the Islamist

terrorist group ISIS in Paris. Her father claims in the suit that Google, which owns the video platform YouTube, should be held responsible for permitting the platform's recommender algorithms to promote content that radicalized the terrorists who killed Gonzalez. The final ruling, expected by the end of June, could make digital platforms liable for their algorithms' recommendations.

Whatever decision the court makes, the case highlights an urgent question: how can societies govern adaptive algorithms that continually change in response to people's behaviour? YouTube's algorithms, which recommend videos through the actions of billions of users, could have shown viewers terrorist videos on the basis of a combination of people's past behaviour, overlapping viewing patterns and popularity trends. Years of peer-reviewed research shows that algorithms used by YouTube and other platforms have recommended problematic content to users even if they never sought it out[1]. Technologists struggle to prevent this.

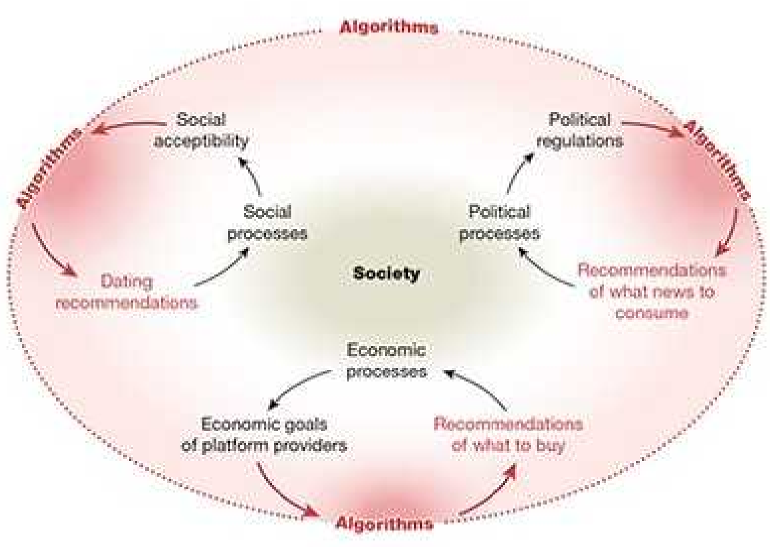
Adaptive algorithms that react to collective action have been linked to other wide-ranging effects. During the 2010 stock market crash, adaptive algorithms responded to a mass stock sell-off initiated by humans and helped it to spiral out of control[2]. An adaptive algorithm that recommended new connections on the professional social network LinkedIn played a part in helping hundreds of thousands of people to move jobs[3]. Generative artificial-intelligence (AI) programs in chatbots have adapted to human behaviour in ways that include learning to parrot racist hate speech. As AI systems are deployed even more widely, the lives of people and adaptive algorithms will become increasingly intertwined.

To regulate such algorithms and ensure a safe, beneficial role for them in society, researchers must build knowledge in science and engineering on collective patterns of human–algorithm behaviour. They need to classify, explain and forecast these patterns, as well as develop interventions that can prevent problems. This will require academics to break down disciplinary boundaries, develop new methods and work collaboratively with communities affected by algorithms. Here, I lay out four steps to develop this science.

# A complex system

An adaptive algorithm is any software that regularly updates its own behaviour based on outcomes that it also influences. The underlying mathematical principles of such algorithms can vary, from simple correlation to deep learning[4]. Many systems are ensembles of multiple models combined into a unified output, such as a video suggestion or a stock trade. When humans respond to those outputs, adaptive algorithms update their own behaviour in turn.

The resulting feedback loop of human–algorithm behaviour is a complex system, in which multiple actors adapt in response to each other and small changes in one part of the system, such as an algorithm adjustment, can lead to unexpected consequences or even collapse. When YouTube's algorithms suggest terrorist recruitment videos, they are responding to collective audience behaviour, as well as campaigns from terrorist groups and counter-campaigns from governments and civil society[1,5].




Measuring algorithmically infused societies

Few patterns of human–algorithm behaviour are exclusively good or bad for society. When YouTube connected some viewers to content posted by members of terrorist groups, its algorithms might have contributed to

clustering — a term for identifying and connecting people who have similar tastes in news or culture, whether or not they previously viewed such videos[6]. These links were suggested to a small proportion of the platform's users at the time — more than one billion in 2015. Similar clustering patterns have brought together influential social movements, given people access to new employment networks and connected people with similar medical conditions to organize the search for cures. A blanket ban on clustering algorithms could reduce these public goods and threaten basic human rights, such as the freedom of association.

This scientific and moral thicket makes it difficult for technologists to safety-test algorithms before their release, to forecast problems, to predict the effects of software changes or to achieve a stable state[7]. The scientific uncertainty has also created questions about who bears responsibility. "It takes two to tango," wrote Nick Clegg, president of global affairs at the social-media company Meta, arguing that because the company's algorithms respond to unpredictable human behaviour, companies should be held less responsible for the outcomes. Industry advocates — those not directly involved in the case but who have a strong interest in its outcome — have filed many 'amicus briefs' in *Gonzalez* v. *Google*, arguing a similar point. In oral arguments in the case, Google's legal counsel argued that the company should legally be protected from liability whether or not its recommender algorithms were biased in favour of videos by terrorist recruiters (see go.nature.com/3hvcimz).

Scientists and technology firms have built relationships to study algorithms and society[8]. In 2018, for example, Facebook (now known as Meta) established initiatives, such as Social Science One, to share data with academic researchers studying the impact of the platform on democratic elections. But academics have also raised concerns that firms sometimes obstruct research. In 2021, Meta cut off data access to researchers at New York University after the academics discovered systematic gaps in transparency data that Meta was providing to academics and regulators about posts before and after the 6 January US Capitol Hill riots. Meta said that it fixed the error. As co-founder of the Coalition for Independent Technology Research, a non-profit organization, I help to defend the right to study the impact of technology on society.

Because companies struggle to control algorithms directly, governments have pressured tech firms to police human behaviour. In 2021, the European Union introduced regulation requiring companies to monitor and remove terrorist content within one hour. Social-media platforms now employ large teams of workers to monitor the activities of billions of people to stop algorithmic harms by moderating problematic content[9].

YouTube says that violent extremist groups and criminal organizations are prohibited from using the platform, and that it is committed to removing this content (see go.nature.com/44aakry). It says that YouTube recommendations generally reflect users' personal preferences and online habits, and that recommendations actively discourage users from viewing radicalizing content by favouring mainstream, authoritative sources.

Policing everyone, everywhere, all at once is costly, laborious and error-prone. Content moderators make frequent mistakes that restrict legitimate rights, while corporate moderation policies are opaque and rarely open to appeal[9]. Furthermore, there is no independently verifiable evidence to support corporate claims that this digital policing keeps anyone safer. But policing people rather than algorithms will remain the status quo in algorithm governance — until advances in science enable reliable, verifiable interventions to be found.

## Classify patterns of behaviour

Because adaptive algorithms are so recent, researchers and policymakers lack clear, shared language to describe how they interact with humans. Just as meteorologists classify different kinds of storm, researchers need to reliably detect and describe patterns of behaviour involving algorithms and humans in ways that scientists and the public can understand.



Relatives of Nohemi Gonzalez, who died in the 2015 Paris terror attacks, have begun a US Supreme Court case against YouTube's owner Google.Credit: Shawn Thew/EPA-EFE/Shutterstock

Decades of interpretive research from journalists, activists and academics has started to describe recurring algorithm-involved incidents[2] but they lack the precision required. For example, polarization, radicalization and organization of social movements all involve clustering, but a more detailed classification system would identify both the algorithm designs and the social forces involved — such as homophily, a human tendency to associate with people who are similar to ourselves. Classifications would also differentiate clustering from algorithm-involved patterns that drive people apart. Because harms involving algorithms are often noticed first by journalists and members of the public, these groups need to be included in such research.

To build this classification system, researchers need a shared resource of cases and evidence. Technologists have already begun to collect cases in the AI Incident Database, a log of harms or near harms involving AI systems,

based mostly on journalism. But companies worried about lawsuits might refuse to provide useful forensic information when problems occur. Researchers therefore need to work together with the media, affected communities and algorithm designers to collect more systematic and detailed data on harmful and beneficial incidents alike.

## Explain how human–algorithm behaviour emerges

Governance of algorithms will require reliable explanations for how things could go wrong. But neither technologists nor social scientists have the language or models to create these explanations.

To a technologist, an explanation is an account of why an algorithm took a specific action, compared with what it might have done otherwise. In the case of YouTube, computer scientists might be able to explain what part of a mathematical matrix contributed most to the algorithm's recommendation of a terrorist video at one moment. But they aren't easily able to suggest changes that would have prevented terrorist recruitment, because that would require a better understanding of human behaviour. That's why computer scientists cannot say in advance whether a given design will become more harmful or less harmful over time[10].

Humanities researchers and social scientists, meanwhile, are working to develop algorithm-informed explanations of collective behaviour. A 2021 study of Twitter users showed that posts expressing moral outrage received more likes than did posts showing other emotions, and that this feedback caused people to express more outrage in the future[11]. But studies such as this can't currently explain how algorithms respond or what they contribute to these dynamics[7].



[Everyone should decide how their digital data are used — not just tech companies]

One way forward comes from outside academia. When lawyers for the Gonzalez family appealed to the Supreme Court to look into the daughter's death, the most compelling evidence on the outcomes of YouTube's algorithm came from a crowdsourced investigation called YouTube Regrets. In the project, run by the non-profit Mozilla Foundation in San Francisco, California, and unrelated to the Paris attacks, more than 37,000 volunteers monitored what YouTube suggested to them and flagged the videos that they regretted having seen in their feed for being inaccurate, offensive and violent. By combining volunteers' reports, researchers were able to show that the platform was recommending thousands of videos that viewers considered problematic[12]. The project did not look specifically at content on terrorism. YouTube says it offers viewers control over their recommendations, including the ability to block videos or channels, and that it welcomes research on its platform.

Because many adaptive algorithms present a different experience to each person, research should study those interactions en masse. In 2022, Northeastern University in Boston, Massachusetts, launched the National Internet Observatory to record and study human and algorithm behaviour in everyday life. The observatory, which has received US$15.7 million in funding from the US National Science Foundation, is recruiting tens of

thousands of people to install software that monitors what algorithms are doing, and securely archives those records with people's consent. Participants will also answer regular survey questions about their lives. As one of the observatory's science advisers, I think it will provide crucial data across disciplines to explain how and why algorithms behave the way they do.

More collective endeavours such as this are needed. Researchers need collaborations, observatories and experiments that combine diverse human experiences with computational methods. Academics and communities need to collaborate to identify circumstances in which runaway, damaging feedback loops arise. Funders need to support such interdisciplinary studies — and journals need to publish them.

## Reliably forecast and intervene

Progress towards flourishing digital environments will ultimately depend on successfully providing benefits and preventing harms. This has two parts: forecasting what could happen, and intervening effectively.

Technologists and economists have already developed software to simulate interactions between populations of people and recommendation algorithms. Simulations could some day enable designers to explore possible outcomes of changes in algorithm design[13]. These simulations could be improved with citizen science information from projects such as YouTube Regrets and the National Internet Observatory. As with weather models, these simulations would ultimately be evaluated on the basis of how well they forecast events outside the laboratory.

Researchers can also forecast collective harms by monitoring algorithms in real time. Over the past few years, my lab has worked together with online communities of millions of people to identify patterns in algorithm behaviour that occur before, and might predict, incoming waves of harassment. We are now analysing results from an experiment that uses these forecasts to allocate messages that studies have shown can reduce and prevent online harassment[14].



The powers and perils of using digital data to understand human behaviour

Although forecasts might help people to avoid problems, the most useful research will identify effective, reliable and reproducible interventions to prevent harm. How should technologists redesign their software? What effects can we expect from certain policies? And can the public confidently organize for change? In 2022, YouTube Regrets ran a randomized trial, involving more than 20,000 people, to test the effects of a new ‘Stop recommending’ button on the platform. Using machine learning to analyse 500 million video recommendations, the group found that YouTube’s controls reduced unwanted recommendations by 11–43%, but that the platform kept recommending substantial quantities of unwanted content anyway[15].

To develop the science of algorithm and human behaviour, the diversity of who does research must broaden. Because adaptive algorithms behave differently for different people, researchers in dominant cultures are often unaware of harms that affect marginalized communities. In 2021, only 19 computer-science PhDs were awarded to Black or African American scholars in the United States and Canada, out of 1,493 PhDs in total. Only 24 identified as Hispanic like me. These monocultures help to explain why activists, journalists and interpretive scholars (such as ethnographers) in more diverse fields have been so effective at observing harms that technologists failed to anticipate. Scientific funders should prioritize

community and citizen-science projects that involve the public. Research collaborations should include disciplines in the social sciences and humanities that have developed broad understandings of diverse human experience. In the long term, researchers must work to diversify their institutions.

## Cultivate hybrids of science and engineering

The biggest question facing the science of human–algorithm behaviour is whether science can be done at all. Critics might argue that establishing a reliable, replicable knowledge base is not possible in a system that responds differently in new situations.

To many computer scientists, human behaviour is too unpredictable to provide reliable guidance on whether interventions will work[10]. Social scientists worry that, because software code is frequently changed, new discoveries could be invalidated as soon as companies change an algorithm's underlying code[2].

Ironically, the persistence of algorithmic harms is the best reason to hope that they can be understood. A 2022 study of six of the most popular search engines (Google, Bing, Yahoo, Baidu, Yandex and DuckDuckGo) found that in searches for images of migrant population groups, all six tended to amplify prejudiced and dehumanizing material[16]. One reason algorithms are curiously predictable is that their creators are so similar. Algorithm makers operate in similar legal environments, face similar economic conditions, receive similar training and use similar data sets[17]. The recurrence of algorithmic harms across many systems and situations could make them classifiable and preventable.

Companies have strong incentives to resist studies that could suggest they are responsible for harms involving their products. That's why future studies will depend on innovations in adversarial data access, ethics and privacy. Researchers should prioritize methods that don't rely on companies, while supporting laws to require algorithm makers to collaborate with researchers. The European Union has recently agreed to mandate research access to tech-

company data (see go.nature.com/4kvjdrd), and US lawmakers are considering the same.

## Next steps

Creating this science of human–algorithm behaviour is a 10- to 20-year challenge. Researchers should form a long-term community of enquiry that spans disciplines and involves ongoing participation from the public. To start, the community should identify and propose initiatives for information-sharing and collaboration, both with and without participation from companies. It should also collectively envision and manage the creation of privacy-protecting research infrastructures. These scientific instruments could integrate work on classifying and explaining human–algorithm behaviour with innovations in forecasting and testing interventions.

I think that adaptive algorithms can deliver widespread benefits to society, and that their harms will some day become manageable. That can only happen if algorithms are designed and governed with reliable research. A science of human–algorithm behaviour can't bring Nohemi Gonzalez back, but whatever the Supreme Court decides later this year, the Gonzalez family's case against tech firms should spur science and society to act.

*Nature* **617**, 248-251 (2023)

*doi: https://doi.org/10.1038/d41586-023-01521-z*

## References

1. Yesilada, M. & Lewandowsky, S. *Internet Pol. Rev.* **11**, 1652 (2022).
2. Matias, N. & Wright, L. *Just Tech* https://doi.org/10.35650/JT.3028.d.2022 (2022).
3. Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E. & Aral, S. *Science* **377**, 1304–1310 (2022).

4. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: an Introduction* 2nd edn (MIT Press, 2018).

5. Fernandez, A. M. *Here to Stay and Growing: Combating ISIS Propaganda Networks* (Brookings Institution, 2015).

6. Murthy, D. *Am. Behav. Sci.* **65**, 800–824 (2021).

7. Bak-Coleman, J. B. *et al. Proc. Natl Acad. Sci. USA* **118**, e2025764118 (2021).

8. Whittaker, M. *Interactions* **28**, 50–55 (2021).

9. Gillespie, T. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (Yale Univ. Press, 2018).

10. Ekstrand, M. D., Das, A., Burke, R. & Diaz, F. *Found. Trends Inf. Retrieval* **16**, 1–177 (2022).

11. Brady, W. J., McLoughlin, K., Doan, T. N. & Crockett, M. J. *Sci. Adv.* **7**, eabe5641 (2021).

12. McCrosky, J. & Geurkink, B. *YouTube Regrets* (Mozilla Foundation, 2021).

13. Stavinova, E. *et al.* Preprint at https://arxiv.org/abs/2206.11338 (2022).

14. Matias, J. N. *Proc. Natl Acad. Sci. USA* **116**, 9785–9789 (2019).

15. Ricks, B. & McCrosky, J. *Does This Button Work? Investigating YouTube's Ineffective User Controls* (Mozilla Foundation, 2022).

16. Urman, A., Makhortykh, M. & Ulloa, R. *Hum. Soc. Sci. Commun.* **9**, 130 (2022).

17. Caplan, R. & Boyd, D. *Big Data Soc.* **5**, 2053951718757253 (2018).

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01521-z

# Work

- **Time tracking for early-career researchers: a practical guide** [08 May 2023]
  • Time management isn't taught at graduate school, but a better understanding of how you pass the hours can help you achieve your goals.
- **Mimicking of human skin to build wearable sensors** [08 May 2023]
  • Yichen Cai creates thin, flexible devices that could have myriad uses, from wearable blood-pressure monitors to touch sensors for robots.

- CAREER FEATURE
- 08 May 2023

# Time tracking for early-career researchers: a practical guide

Time management isn’t taught at graduate school, but a better understanding of how you pass the hours can help you achieve your goals.

- Linda Nordling

Find a new job



Keeping track of time is key not only to working efficiently but also to maintaining a healthy work–life balance.Credit: Mariam Al kassar/Shutterstock

For his PhD in drug design and development, Byron Mubaiwa found planning and executing his laboratory work a breeze. “I spent 10 hours in the lab a day, 5 days a week,” the South African says. “I had a lab notebook I completed every day, and I had a research plan, so I knew exactly what I needed to do, and when.”

But as Mubaiwa moved on to writing his doctoral thesis, managing his performance became a lot more complicated. It had been easy to focus in the lab, but he would now be writing off-campus while juggling a full-time job working as a warehouse pharmacist. How to accomplish that was something he had to work it out himself.

“Nobody spoke to me about time management,” says Mubaiwa, who gained his PhD this year and now works as a project manager for TASK, a clinical-research institute based in Cape Town, South Africa. He started by blocking out times in his Google Calendar to work on his thesis, then tried different time-tracking apps, which helped him to focus.

Then, “everything changed” when he tried an app called Jira — a tool developed by the software company Atlassian for collaboration across teams, but which he found ideal for his thesis writing. “It allowed me to put timelines of when I wanted to complete each chapter and allowed me to estimate timelines and log actual time spent against the estimate.” The app enabled him to generate reports detailing his performance, which led to his accountability levels increasing and his productivity improving “exponentially”.

Time tracking, he found, was a great motivator, as well as a way to organize his work, and it became a skill that he now uses every day while managing complex clinical-research projects for TASK. “Nowadays, tracking is a part of me — I track everything,” he says.

## Keep it simple

Ask any doctoral candidate what they find challenging about their work, and time management tends to be up there. In a 2022 survey of more than 3,200 graduate students around the world (*Nature* **610**, 805–808; 2022), half

worried that they would not finish their degrees on time and more than two-thirds reported struggles with work–life balance.



[How to use ‘20% time’ to manage pet projects]

Yet how to manage time is rarely something that universities rarely teach young academics. Someone who helps researchers with this aspect of their work is Melanie Smith, a former law lecturer who now works as an academic coach, offering writing support for researchers at all stages of their careers.

“Time tracking is one of the first things I ask clients to start doing,” says Smith, who spent more than 15 years teaching law at UK universities, in Manchester and Cardiff, and now lives in Ireland. She suggests using something like myhours.com, an app that is basic, but easy and quick to use.

Many of her clients are hesitant at first, saying that they don’t have time to deal with another piece of software. But Smith thinks there might be deeper, more personal reasons why they resist.

“I have people come to me and say: ‘I’ve been working on this paper 8 hours a day and I’m just not making any progress.’ But with time tracking, they find that what they thought was 8 hours was really 45 minutes. The rest of the time was them sitting feeling emotional and writing e-mails, going on

the Internet or pretending to do research," she says. "Lots of time tracking is about breaking that delusion."

But there is more to successful time tracking than just noting down what was done when. No matter how meticulously those things are recorded, the real benefit comes when the data are used to plan future activities, says Smith.

Mubaiwa says that he used time tracking not just to analyse his progress, but also to keep track of things that blocked his productivity. Gradually, he was able to schedule some slack time into his schedule to allow for the distractions that he knew would inevitably creep in.

He also learnt how to adapt when things took more or less time than he expected. "As long as I felt organized, both hitting the deadlines and being late for them felt like part of the plan," he says.

But he has also realized that his methods won't work for everyone. "You have to find what works for you. Keep an open mind about the various techniques out there, but ultimately be kind to yourself."



Sijin Chen schedules timely breaks to ensure that she's neither too tired nor too wired to relax.Credit: Sijin Chen

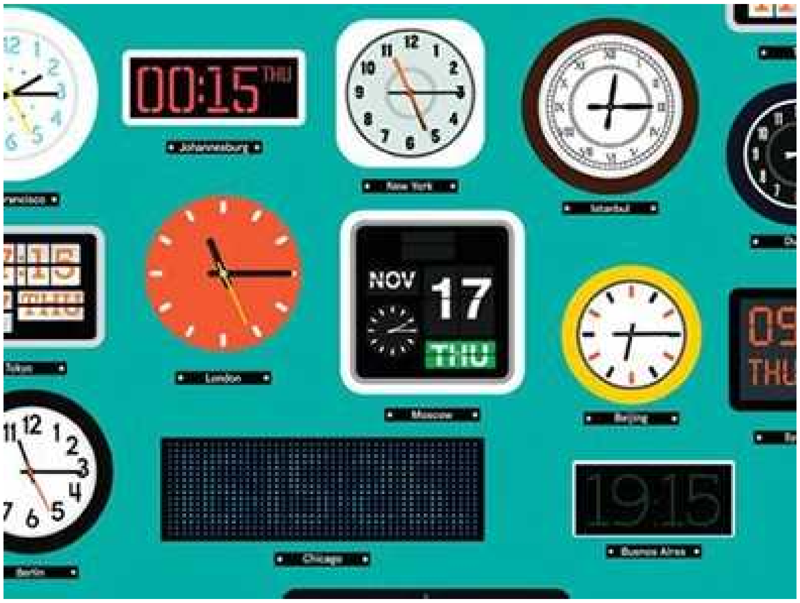
For some, too much tracking can be bad. Sijin Chen, a second-year PhD candidate at the National University of Singapore, tried to track every single hour of her day — whether she was working, relaxing or sleeping. But it just made her paranoid, she says.

Chen, who is from China, is studying consumer behaviour for her doctorate. She says she is "infatuated" with productivity and describes herself as a "small-time influencer". Her videos on Bilibili, a Chinese variation on YouTube, in which she shares productivity tips and insights from her research, get an average of 50,000 views within a week of posting.

"People in China are starting to look out for work–life balance, especially the younger generation," she says. However, for Chen, that balance has been elusive. "I wouldn't say I'm good at time management by nature," she says. "I'm very easily distracted."

## Manage energy

To maximize her productivity, Chen began scheduling everything, including her downtime, but found it to be counterproductive. Relaxing on demand proved difficult. And even if she took herself out of her working environment — by going to an art gallery, for instance — she struggled to immerse herself in the experience. "I couldn't stop myself from thinking about work and my endless to-do lists. I was anxious, and then I got anxious over being anxious."




[Time management for scientists]

What she does now is schedule blocks of time for work that are short enough to ensure that she does not go into her scheduled downtime either too tired or too wired to relax: “I schedule breaks before I reach that point of exhaustion when all you can do is scrolling on your phone, which is not actually relaxing.”

According to a 2021 survey, 87% of UK university staff said that their workload had increased in size or intensity over the previous three years, and one in ten academics described their workload as “entirely unmanageable” — nearly twice as many as in 2016 (see go.nature.com/3pcg9wd).

Academics who know how they spend their time can use that knowledge to push back when faced with unreasonable demands, says Smith. “When people come and ask you to do yet another thing, you can whip out your chart and say: ‘Hey, listen, this is how I spend my time already. Which part of this do you want me to take away?’”

## Know thyself

For those wrestling with time management, Mubaiwa has some tips. He suggests that doctoral students run their PhDs as a series of mini-projects, mapping out high-level timelines, then breaking them into subtasks with detailed timelines for each.

If things become difficult, he suggests staying the course and first changing small things before throwing out the whole plan. Then, most importantly, he says: “Learn about yourself — understand your personality and try to align who you are with how you get things done.”

Smith says that time tracking is a way of bringing accountability to the practice of scheduling your time, to make sure that you achieve your goals. “It tells you what you are actually doing with your time, not what you imagine you are doing with it.”

*Nature* **617**, 427-428 (2023)

*doi: https://doi.org/10.1038/d41586-023-01552-6*

---

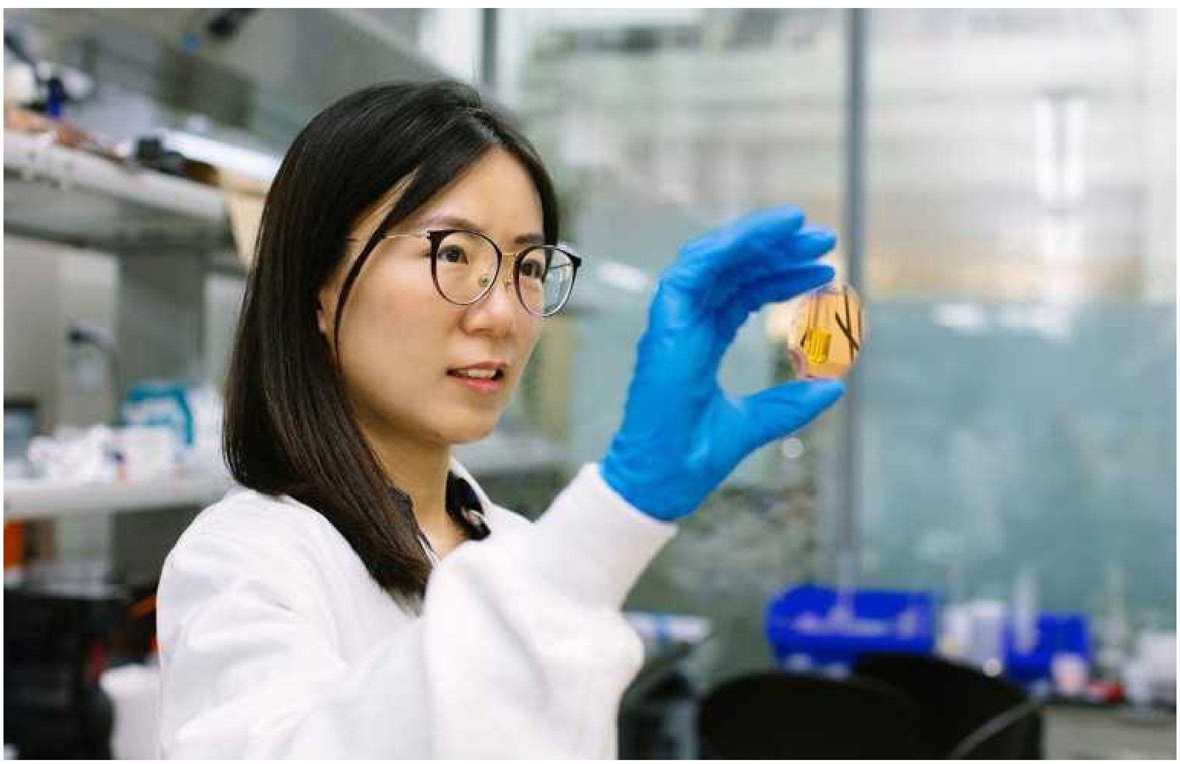
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01552-6

- WHERE I WORK
- 08 May 2023

# Mimicking of human skin to build wearable sensors

Yichen Cai creates thin, flexible devices that could have myriad uses, from wearable blood-pressure monitors to touch sensors for robots.

- Linda Nordling



Yichen Cai is a materials scientist at the King Abdullah University of Science and Technology in Thuwal, Saudi Arabia. Credit: Anastasia Serin/KAUST

I am a materials scientist at the King Abdullah University of Science and Technology (KAUST) in Thuwal, Saudi Arabia, and my work involves creating thin, flexible, conductive films that mimic human skin. This ‘e-skin’ has many potential applications, especially in health care, where it can be used to build wearable sensors to measure things such as blood pressure, movement and body temperature. It could also allow robots to sense touch, like humans do, or deliver drugs into the body.

KAUST is the first mixed-gender university in Saudi Arabia. Here, female employees have work and education opportunities equal to those of male employees.

In my laboratory, I make films as thin as one nanometre. In this photo, I’m standing in front of our chemical-vapour deposition furnace, where my team and I build the films. We do this on a very flat substrate: the disc I’m holding here. In the reflection on it you can see a vial filled with a fluid. This is the catalyst that we use to make the films out of small monomers, which bond in the furnace to form polymers. It took us three years to perfect our synthesis process.

In the furnace, I can vary the films’ pore structure and thickness, which affects their electrical and mechanical properties. This is directly related to the sensitivity and elasticity of the e-skin. We are still in the early stages of our research. One challenge is to scale up what we do in the lab to something commercially viable. We also need to work on the stability of our e-skin. We have only tested it for short-term use, but some applications might require sensors that work for days, or even months.

Thin, flexible sensors could be used in industrial settings as well: to detect toxic leaks rapidly and reliably, for example. My colleagues and I are also working on developing a film that can remove salt from seawater and brine, as an alternative to existing desalination methods. The opportunities are endless.

*Nature* **617**, 430 (2023)

*doi: https://doi.org/10.1038/d41586-023-01553-5*

---

This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01553-5

# Research

- **Human pangenome supports analysis of complex genomic regions** [10 May 2023]
  • A pangenome is a collection of DNA sequences that reveals genetic variation between individuals. Four scientists discuss the generation of a human pangenome, and what insights can be gained from it.
- **Loophole-free Bell inequality violation with superconducting circuits** [10 May 2023]
  • A loophole-free violation of Bell's inequality with superconducting circuits shows that non-locality is a viable new resource in quantum information technology realized with superconducting circuits, promising many potential applications.
- **High-throughput printing of combinatorial materials from aerosols** [10 May 2023]
  • The authors report a high-throughput combinatorial printing method capable of fabricating materials with compositional gradients at microscale spatial resolution, demonstrating a variety of high-throughput printing strategies and applications in combinatorial doping, functional grading and chemical reaction.
- **Near-frictionless ion transport within triazine framework membranes** [26 Apr 2023]
  • The authors develop a strategy that allows the diffusion limit of ions in water to be approached for large-area, free-standing, synthetic membranes using covalently bonded polymer frameworks with rigidity-confined ion channels.
- **A draft human pangenome reference** [10 May 2023]
  • An initial draft of the human pangenome is presented and made publicly available by the Human Pangenome Reference Consortium; the draft contains 94 de novo haplotype assemblies from 47 ancestrally diverse individuals.
- **Increased mutation and gene conversion within human segmental duplications** [10 May 2023]
  • A study comparing the pattern of single-nucleotide variation between unique and duplicated regions of the human genome shows that mutation rate and interlocus gene conversion are elevated in duplicated regions.
- **Recombination between heterologous human acrocentric chromosomes** [10 May 2023]

• Comparisons within the human pangenome establish that homologous regions on short arms of heterologous human acrocentric chromosomes actively recombine, leading to the high rate of Robertsonian translocation breakpoints in these regions.

- ## COVID-19 amplified racial disparities in the US criminal legal system [19 Apr 2023]
  • A study shows that, although the number of incarcerated people in the USA decreased during the first year of the COVID-19 pandemic, the fraction of incarcerated Black and Latino individuals increased.

- ## A somato-cognitive action network alternates with effector regions in motor cortex [19 Apr 2023]
  • Functional MRI studies across ages show that the classic homunculus of the motor cortex in humans is in fact discontinuous, alternating with action control-linked regions termed the somato-cognitive action network.

- ## Learnable latent embeddings for joint behavioural and neural analysis [03 May 2023]
  • A new encoding method, CEBRA, jointly uses behavioural and neural data in a (supervised) hypothesis- or (self-supervised) discovery-driven manner to produce both consistent and high-performance latent spaces.

- ## A druggable copper-signalling pathway that drives inflammation [26 Apr 2023]
  • Cellular uptake of copper(ii) by CD44 has a key role in regulating cellular plasticity via copper(ii)-dependent downstream signalling events.

- ## Structural atlas of a human gut crassvirus [03 May 2023]
  • A cryo-electron microscopy reconstruction of the virus ΦcrAss001 provides insights into the functions of the viral gene products in capsid assembly and infection.

- ## Ligand and G-protein selectivity in the κ-opioid receptor [03 May 2023]
  • Active-state structures of the κ-opioid receptor in complexes with the G-protein heterotrimers Gi1, GoA, Gz and Gg provide insights into the actions of hallucinogenic opioids and G-protein-coupling specificity at the κ-opioid receptor.

---

- NEWS & VIEWS FORUM
- 10 May 2023

# Human pangenome supports analysis of complex genomic regions

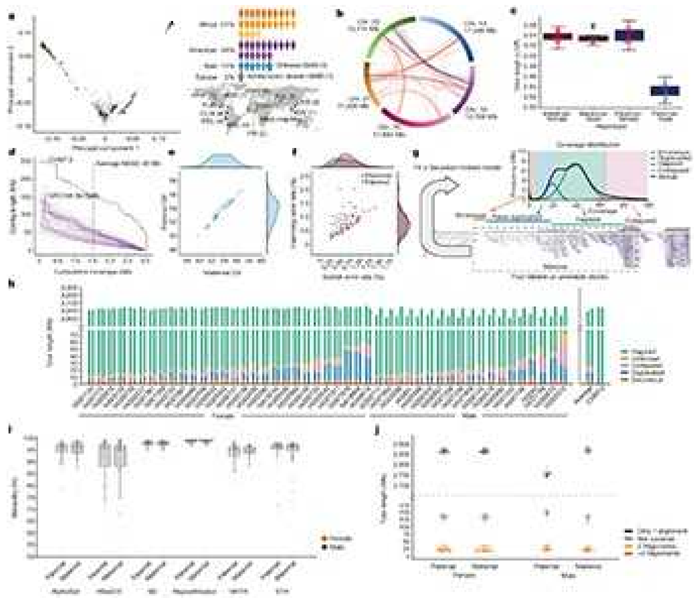
A pangenome is a collection of DNA sequences that reveals genetic variation between individuals. Four scientists discuss the generation of a human pangenome, and what insights can be gained from it.

- Arya Massarat ,
- Melissa Gymrek ,
- Brian McStay &
- …
- Hákon Jónsson

Show authors

## ARYA MASSARAT & MELISSA GYMREK: Describing genetic diversity with graphs

Reference genomes are crucial coordinate systems for genomic analyses. However, the references that scientists currently work from when studying humans (the draft human genome[1] and its complete, gap-free successor[2], dubbed T2T-CHM13) are both based mostly on single individual genomes. A linear genome sequence of this type cannot adequately represent genetic diversity within our species. Instead, such diversity is more accurately described using a graph-based system of branching and merging paths. In a paper in *Nature*, Liao *et al.*[3] describe the first human reference pangenome — a collection of genome sequences compiled into a single data structure.



Read the paper: A draft human pangenome reference

The use of human reference genomes from single individuals is problematic because it introduces biases in how sequences from other human genomes are interpreted. For instance, sequences from other genomes are first commonly aligned to the reference (read mapping) and then reduced to a set of differences from that reference (variant calling). Both processes might yield different results if a different person's DNA had been used to generate the original reference. This is particularly true for highly diverse and structurally complex regions of the genome. Furthermore, there are hundreds of megabases of DNA that cannot be captured in a reference based on a single genome, because they exist in only a subset of humans[4,5].

A pangenome representing many genomes from different ancestries could overcome these issues. However, constructing a pangenome is a complex task. Breakthroughs in the past decade in long-read sequencing technology and computational methods have now enabled this vision to be realized.

Liao and colleagues first generated 94 genome assemblies from 47 individuals (one for each of the two sets of chromosomes that each individual carries). The individuals represent diverse ancestries from around the globe. The assembled genomes, which were generated using a combination of long-read and other sequencing technologies, are highly

accurate and nearly complete, and include 119 million base pairs of sequence not included in the draft human reference genome.

The authors used three graph-building methods to construct pangenomes from these assemblies. One of these methods aligns all sequences simultaneously; the others use one genome as a reference and align each subsequent sequence iteratively. The result is a set of publicly available pangenome graphs, along with a rich ecosystem of open-source tools and standardized file formats that researchers can use in a similar way to a linear reference genome.

Liao *et al.* demonstrated that using their pangenomes for read mapping and variant calling resulted in 34% fewer errors in calling small variants (those shorter than 50 bases) than did using a linear reference. The difference was particularly pronounced in challenging repetitive DNA regions. Impressively, the pangenomes enabled the authors to identify twice as many large genomic alterations, called structural variants, per person than is possible using a linear reference (Fig. 1).

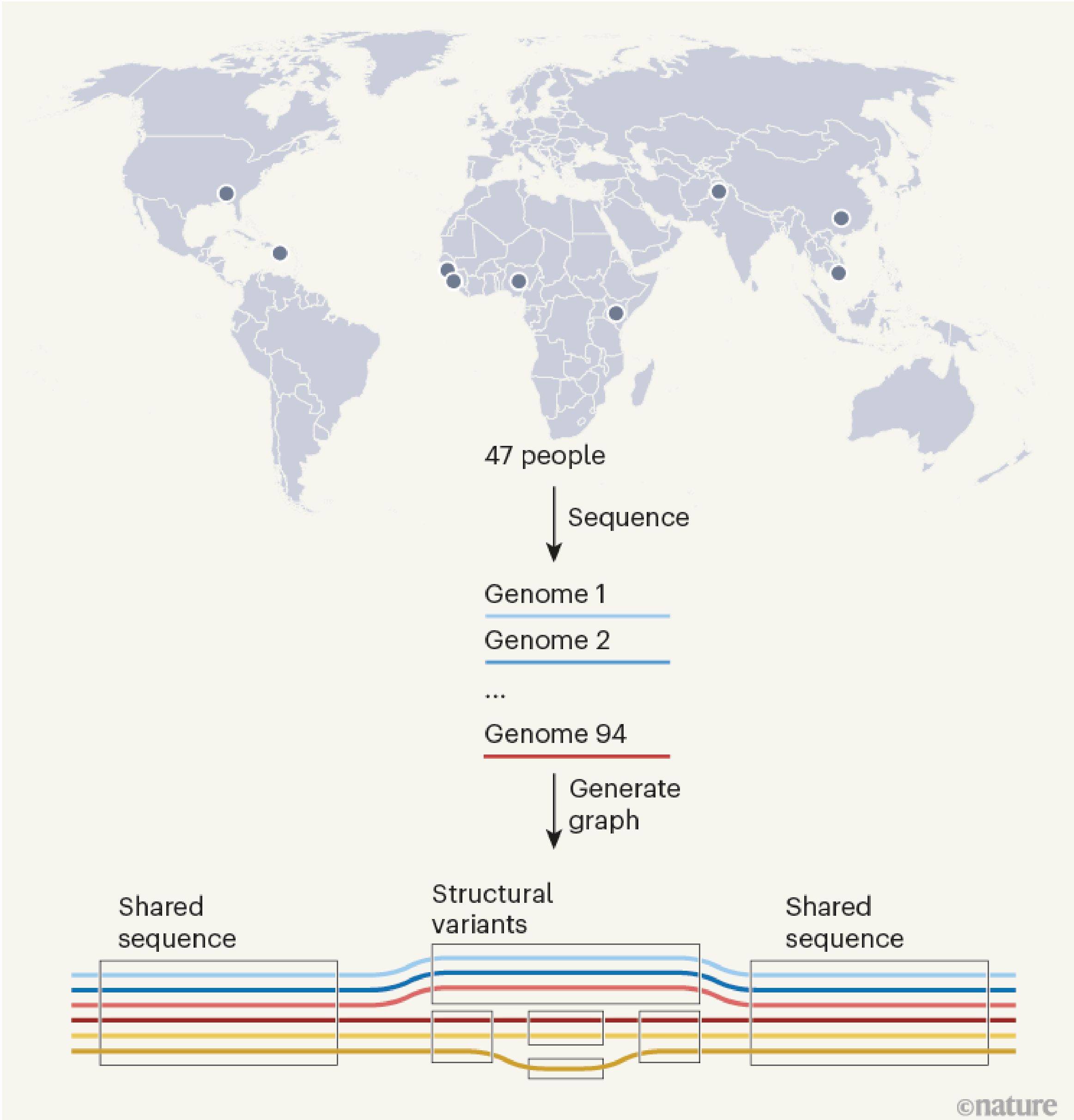


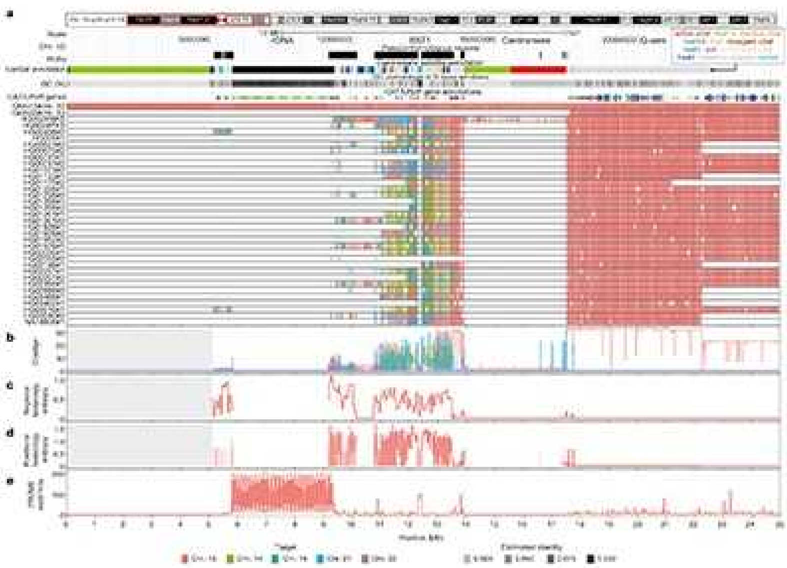

**Figure 1 | A human pangenome.** The genomes of 47 people of diverse ancestries have been used to generate a draft human pangenome reference[3,6,7]. Two whole-genome sequences were generated from each individual (one for each of their two sets of chromosomes). The 94 sequences were aligned to form a pangenome graph, which is conceptually similar to an underground-railway map. Boxed regions indicate sequences present in one or more genomes that at a given site, with branching paths indicating sequence variation. The graphs reveal large genome alterations called structural variants, and enable easy analysis of how they vary between individuals.

The human pangenome reference represents a milestone in human genetics. However, challenges remain. Alignment of sequences against highly variable repetitive regions in the pangenome could be improved by more-accurate assemblies or new algorithms. More samples from diverse groups are also needed. Finally, widespread adoption of the pangenome by scientists could take time, because new methods supporting pangenome analysis are continually being developed, and scientists will often require training to use them.

Continued improvements in methods for building and using pangenomes will enable researchers to overcome these challenges. Use of pangenomes has the potential to transform human genomics. This will ultimately make it easier to discover genetic variants that mediate physical and clinical traits and — it is to be hoped — will eventually lead to better health outcomes for many people.

## BRIAN MCSTAY & HÁKON JÓNSSON: Untangling repeated sequences

Repetitive DNA regions are challenging to sequence, because it is hard to place them accurately in a genome assembly. These regions include segmental duplications (in which sequences more than one kilobase long are repeated elsewhere in the genome) and the short arms (p-arms) of a subset of chromosomes, dubbed acrocentric chromosomes. Two studies in *Nature* now use Liao and colleagues' pangenomes to systematically explore these regions — Guarracino *et al*.[6] to analyse acrocentric p-arms, and Vollger *et al*.[7] to investigate segmental duplications. Their work provides a glimpse of the insights that can be gained from a pangenome reference.



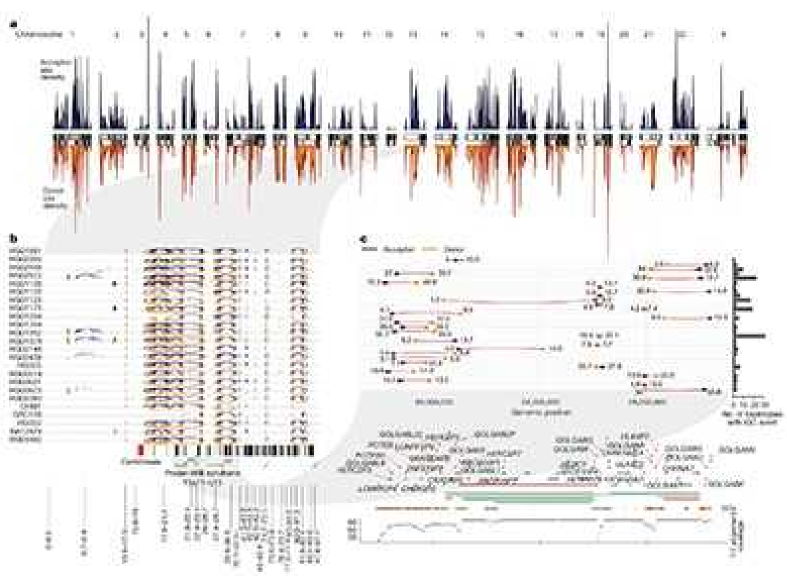
Read the paper: Recombination between heterologous human acrocentric chromosomes

The acrocentric chromosomes (chr13, chr14, chr15, chr21 and chr22 in humans) are those in which the p-arm is considerably shorter than the other (q) arm. Acrocentric p-arms are dedicated to one task: forming sites called nucleoli, where the cell's protein-assembling machines are made[8]. P-arms contain nucleolar organizer regions (which encode the RNAs that drive nucleolar formation), highly repetitive DNA and many other shared sequences. This shared, repetitive DNA reflects a phenomenon called heterologous recombination, whereby different acrocentric p-arms pair and cross over to exchange DNA during the cell divisions that generate sperm and eggs. By contrast, in most chromosomes, pairing and crossover are restricted to two copies of the same chromosome (homologous recombination).

In XY sex chromosomes, which also exhibit heterologous recombination, pairing is aided by short regions of homology (near-identical sequences) shared between X and Y. Guarracino *et al.* constructed a variation graph for acrocentric p-arms using Liao and colleagues' sequences, and found that they contain pseudo-homologous regions (PHRs). Each PHR is a patchwork of sequence blocks that — as the authors discovered when they compared their graphs with T2T-CHM13 — often show more similarity to the other four acrocentric p-arms in T2T-CHM13 than to the T2T-CHM13 version of

themselves. Presumably, these blocks assist heterologous recombination, ensuring that p-arms evolve in concert to preserve their shared role in nucleolar formation.

Robertsonian translocations (ROBs) are phenomena, usually occurring during egg-cell production, whereby the q-arms of two acrocentric chromosomes fuse and most of the p-arms are lost[8]. Guarracino *et al*. identified sequences in PHRs at which the breaks that lead to ROBs occur — indicating that ROBs are collateral damage arising from heterologous recombination. Given that ROBs occur in one in 800 human births, we surmise that heterologous recombination between acrocentric chromosomes is both ongoing and frequent. We expect that, as more genomes are added to the pangenome reference, it will be possible to quantify the frequency of this recombination.



[Read the paper: Increased mutation and gene conversion within human segmental duplications]

Vollger *et al*. used the reference to systematically compare variation in segmental duplications with that in non-repetitive parts of the genome (Fig. 1). They found 60% higher sequence diversity in segmental duplications, and showed that these duplications are highly divergent between populations and individuals.

Genes in segmental duplications are susceptible to interlocus gene conversion (IGC) — an exchange of short DNA sequences between non-homologous parts of the duplicated region. Vollger and colleagues identified IGC events by looking for signs of sequence shuffling in the pangenomes, and concluded that these events are probably one of the main reasons that segmental duplications are so diverse. They found that 799 genes had protein-coding regions affected by an IGC.

It is exciting to see accurate characterization of segmental duplications, because duplicated sequences can fuel the evolution of new, specialized roles for a gene. Vollger *et al*. assessed sequence 'constraint' in duplicated genes, with a particular interest in those duplicated during evolution of the human lineage. Constraint is a measure of sequence variability, with less variation indicating that mutations are detrimental to the organism's viability. Thirty-eight genes were constrained, including members of the *NOTCH2* gene family, which has been linked to human-specific changes in brain size during evolution[9]. The repetitive nature of segmental duplications had previously led to difficulties in assessing constraint for at least 40% of the analysed genes. The authors also found that 171 genes were duplicated and relocated intact to new genomic regions, potentially meaning that their regulation would be rewired. In future, the pangenome project should enable researchers to assess constraint in recently duplicated genes in more depth.

Together, these papers provide a taster of how the human pangenome reference can be used. They reveal how sequence exchanges between repetitive regions of our genome contribute to variation in the population and to our evolution. As the scope of the reference expands, we look forward to further insights into these fascinating genomic regions.

*Nature* **617**, 256-258 (2023)

*doi: https://doi.org/10.1038/d41586-023-01490-3*

# References

1. Lander, E. S. *et al. Nature* **409**, 860–921 (2001).

2. Nurk, S. *et al. Science* **376**, 44–53 (2022).

3. Liao, W.-W. *et al. Nature* **617**, 312–324 (2023).

4. Sherman, R. M. *et al. Nature Genet.* **51**, 30–35 (2019).

5. Ebert, P. *et al. Science* **372**, eabf7117 (2021).

6. Guarracino, A. *et al. Nature* **617**, 335–343 (2023).

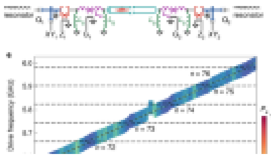
7. Vollger, M. R. *et al. Nature* **617**, 325–334 (2023).

8. McStay, B. *Annu. Rev. Genom. Hum. Genet.* https://doi.org/10.1146/annurev-genom-101122-081642 (2023).

9. Fiddes, I. T. *et al. Cell* **173**, 1356–1369 (2018).

---

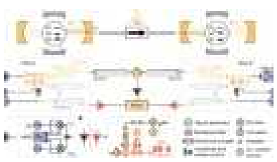
This article was downloaded by **calibre** from https://www.nature.com/articles/d41586-023-01490-3



## Violating Bell’s inequality with remotely connected superconducting qubits

22 April 2019

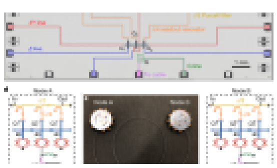
Y. P. Zhong, H.-S. Chang, … A. N. Cleland



## Deterministic quantum state transfer and remote entanglement using microwave photons

13 June 2018

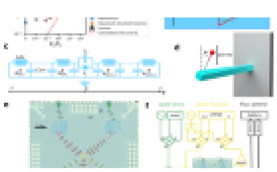
P. Kurpiers, P. Magnard, … A. Wallraff



## Deterministic multi-qubit entanglement in a quantum network

24 February 2021

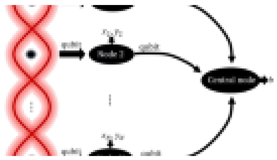
Youpeng Zhong, Hung-Shen Chang, … Andrew N. Cleland



## [Unimon qubit]

12 November 2022

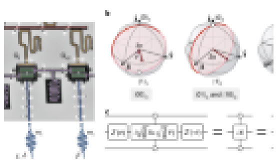
Eric Hyyppä, Suman Kundu, … Mikko Möttönen



## [Demonstrating the power of quantum computers, certification of highly entangled measurements and scalable quantum nonlocality]

22 July 2021

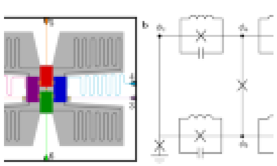
Elisa Bäumer, Nicolas Gisin & Armin Tavakoli



## [High-fidelity three-qubit iToffoli gate for fixed-frequency superconducting qubits]

02 May 2022

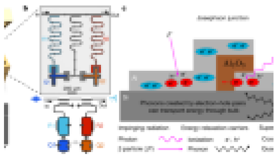
Yosep Kim, Alexis Morvan, … Irfan Siddiqi



## [Realization of efficient quantum gates with a superconducting qubit-qutrit circuit]

16 September 2019

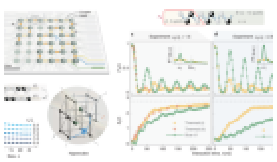
T. Bækkegaard, L. B. Kristensen, … N. T. Zinner



[**Impact of ionizing radiation on superconducting qubit coherence**]

26 August 2020

Antti P. Vepsäläinen, Amir H. Karamlou, … William D. Oliver



**Many-body Hilbert space scarring on a superconducting processor**

13 October 2022

Pengfei Zhang, Hang Dong, … Ying-Cheng Lai

- Article
- Open Access
- Published: 10 May 2023

# Loophole-free Bell inequality violation with superconducting circuits

- Simon Storz ,
- Josua Schär,
- Anatoly Kulikov,
- Paul Magnard,
- Philipp Kurpiers,
- Janis Lütolf,
- Theo Walter,
- Adrian Copetudo,
- Kevin Reuer,
- Abdulkadir Akin,
- Jean-Claude Besse,
- Mihai Gabureac,
- Graham J. Norris,
- Andrés Rosario,
- Ferran Martin,
- José Martinez,
- Waldimar Amaya,
- Morgan W. Mitchell,
- Carlos Abellan,
- Jean-Daniel Bancal,
- Nicolas Sangouard,
- Baptiste Royer,
- Alexandre Blais &
- …
- Andreas Wallraff

Show authors


*Nature* **volume 617**, pages 265–270 (2023)


- 211 Altmetric

- Metrics details


## Abstract

Superposition, entanglement and non-locality constitute fundamental features of quantum physics. The fact that quantum physics does not follow the principle of local causality[1,2,3] can be experimentally demonstrated in Bell tests[4] performed on pairs of spatially separated, entangled quantum systems. Although Bell tests, which are widely regarded as a litmus test of quantum physics, have been explored using a broad range of quantum systems over the past 50 years, only relatively recently have experiments free of so-called loopholes[5] succeeded. Such experiments have been performed with spins in nitrogen–vacancy centres[6], optical photons[7,8,9] and neutral atoms[10]. Here we demonstrate a loophole-free violation of Bell's inequality with superconducting circuits, which are a prime contender for realizing quantum computing technology[11]. To evaluate a Clauser–Horne–Shimony–Holt-type Bell inequality[4], we deterministically entangle a pair of qubits[12] and perform fast and high-fidelity measurements[13] along randomly chosen bases on the qubits connected through a cryogenic link[14] spanning a distance of 30 metres. Evaluating more than 1 million experimental trials, we find an average $S$ value of $2.0747 \pm 0.0033$, violating Bell's inequality with a $P$ value smaller than $10^{-108}$. Our work demonstrates that non-locality is a viable new resource in quantum information technology realized with superconducting circuits with potential applications in quantum communication, quantum computing and fundamental physics[15].

## Main

One of the astounding features of quantum physics is that it contradicts our common intuitive understanding of nature following the principle of local causality[1]. This concept derives from the expectation that the causes of an event are to be found in its neighbourhood (see Supplementary Information section I for a discussion). In 1964, John Stewart Bell proposed an experiment, now known as a Bell test, to empirically demonstrate that theories satisfying the principle of local causality do not describe the properties of a pair of entangled quantum systems[2,3].

In a Bell test[4], two distinct parties A and B each hold one part of an entangled quantum system, for example, one of two qubits. Each party then

chooses one of two possible measurements to perform on their qubit, and records the binary measurement outcome. The parties repeat the process many times to accumulate statistics, and evaluate a Bell inequality[2,4] using the measurement choices and recorded results. Systems governed by local hidden variable models are expected to obey the inequality whereas quantum systems can violate it. The two underlying assumptions in the derivation of Bell's inequality are locality, the concept that the measurement outcome at the location of party A cannot depend on information available at the location of party B and vice versa, and measurement independence, the idea that the choice between the two possible measurements is statistically independent from any hidden variables.

A decade after Bell's proposal, the first pioneering experimental Bell tests were successful[16,17]. However, these early experiments relied on additional assumptions[18], creating loopholes in the conclusions drawn from the experiments. In the following decades, experiments relying on fewer and fewer assumptions were performed[19,20,21], until loophole-free Bell inequality violations, which close all major loopholes simultaneously, were demonstrated in 2015 and the following years[6,7,8,9,10]; see ref. [22] for a discussion.

In the development of quantum information science, it became clear that Bell tests relying on a minimum number of assumptions are not only of interest for testing fundamental physics but also serve as a key resource in quantum information processing protocols. Observing a violation of Bell's inequality indicates that the system possesses non-classical correlations, and asserts that the potentially unknown quantum state has a certain degree of entanglement and purity. This assessment, based on the observed correlations between the chosen input (the choice of measurement basis) and recorded output values (the measurement outcome) of the test, does not rely on knowledge of the inner workings of the system: a property known as device independence[23]. This allows identifying quantum states and measurements[24], certifying the correct functioning of quantum computing devices[25], and establishing common and secret keys between two parties with only limited assumptions about the used devices[26]. Further applications of Bell tests include device-independent randomness generation and expansion, extending a given random bit string in a certified manner[27,28] and

randomness amplification, improving the quality of a source of randomness in a certified manner[29,30], which is a task impossible to achieve by purely classical means.

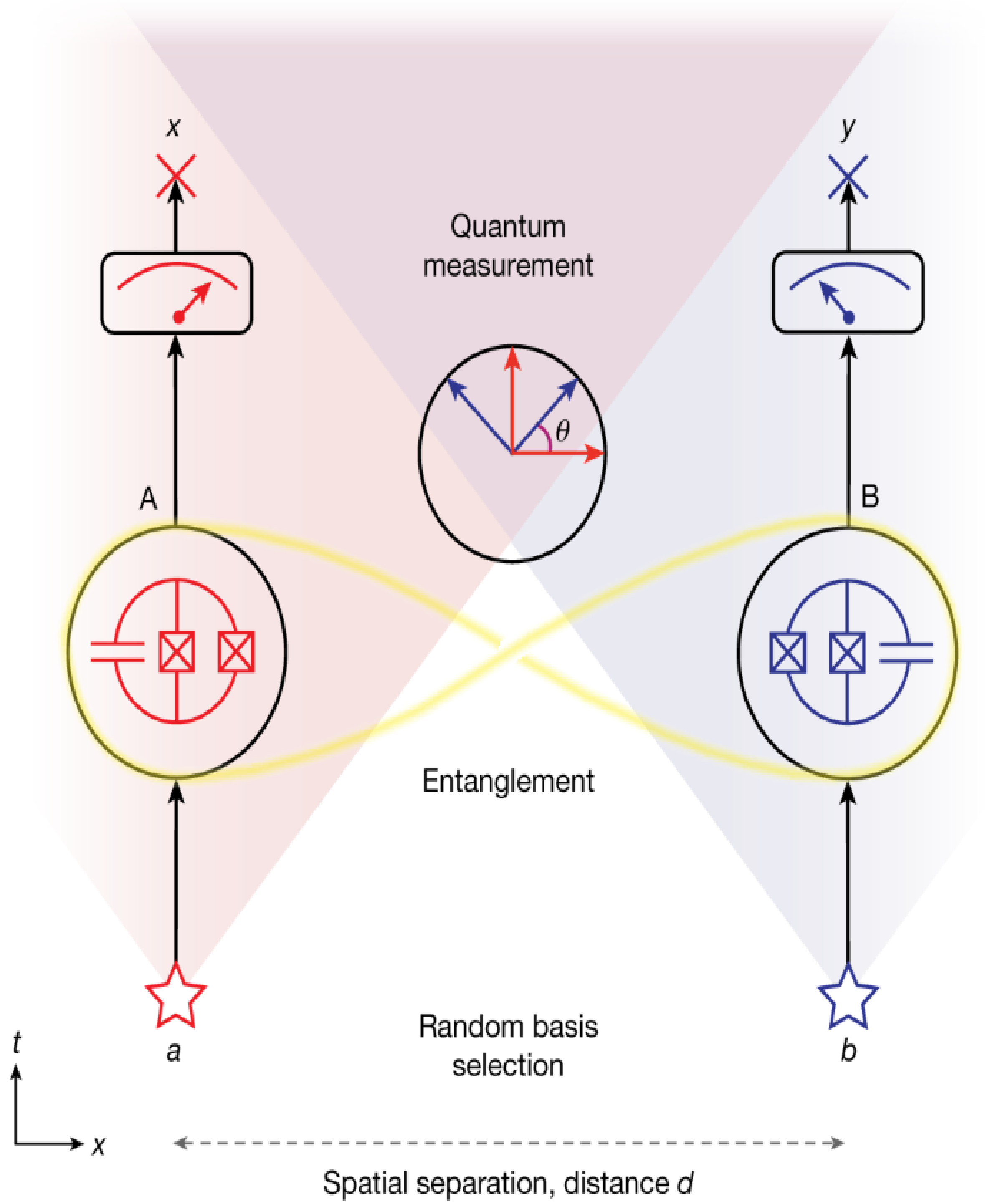
Deploying non-locality as a new resource in the context of superconducting circuits enables new applications in a system that is well set for creating large-scale quantum computers[11,31] and provides quantum communication capabilities. In addition, non-local Bell tests with superconducting circuits are unique as a macroscopic quantum system[32,33,34] is used, which is controlled, entangled and read out exclusively using microwave frequency radiation rather than optical frequency fields.

With superconducting circuits, Bell tests were performed that closed the fair-sampling (or detection) loophole[35], supported the assumption of measurement independence with human choices[36], and used qubits connected by an on-chip 78-cm-long transmission line[37]. Whereas these experiments all relied on additional assumptions, in this work, we set out to demonstrate a loophole-free violation of Bell's inequality using superconducting circuits. The Methods section provides a very brief introduction to the basic properties of superconducting qubits.

Addressing the locality loophole[5] (Supplementary Information section I) in a Bell test with superconducting circuits, typically housed in their individual cryogenic systems, is particularly challenging, as it requires to entangle a pair of qubits located at two sites A and B separated by a large physical distance $d$ with high concurrence ${\mathcal{C}}$ of the entangled state, where ${\mathcal{C}}$ (refs. [38,39]) is a measure of the degree of entanglement present in the system. An individual trial of a Bell test begins at time $t_{\star} = 0$ with the choice of a pair of input bits ($a$, $b$), which determine the basis in which the quantum state of each of the two entangled qubits is read out (Fig. 1). To support the assumption of measurement independence, local basis choices are realized using random number generators (RNGs). If the sites A and B are separated from each other by a sufficiently large distance $d$, the exchange of information between A and B, occurring at most at the speed of light $c$, is prohibited for times $t < t_d = d/c$ according to the laws of special relativity. If the measurement outcomes are obtained during this time interval, the spatial separation thus ensures that the chosen

measurement bases and the corresponding measurement outcomes by the party at one site are unknown to the party at the other site, thereby closing the locality loophole.

**Fig. 1: Schematic of the Bell test experiment.**



Two parties A and B choose random-input bits ($a$, $b$) at the space–time locations indicated by stars and perform measurements on a pair of entangled quantum systems (in this work, superconducting circuit qubits) yielding output bits ($x$, $y$) at space–time locations indicated by crosses. The shaded areas indicate the forwards light cones originating at the space–time location of the random-input-bit-generation events. The inset in the middle indicates the offset angle $\theta$ between the measurement bases of the two qubits (main text).

For each trial of the Bell test, a high-fidelity measurement of the quantum state of the qubits at A and B, which is designed to terminate at time $t < t_d$, is performed. The readout of the qubits results in outcomes $x$ and $y$ taking on values of +1 or −1 if the qubit is detected in the ground \(\left|g\right\rangle \) or excited state \(\left|e\right\rangle \), respectively. Including each and every measurement outcome in the analysis of the Bell test closes the fair-sampling loophole[40,41] (Supplementary Information section I). Furthermore, the memory loophole is closed by statistically analysing the input and output data without assuming that individual trials of the Bell test are independent and identically distributed[18].

To evaluate the result of a Bell test performed in this manner, the averages of the product of the individual measurement outcomes $\langle xy\rangle_{(a,b)}$ at sites A and B are calculated to determine the Clauser–Horne–Shimony–Holt (CHSH) value[4] $S = \langle xy\rangle_{(0,0)} - \langle xy\rangle_{(0,1)} + \langle xy\rangle_{(1,0)} + \langle xy\rangle_{(1,1)}$ given the four possible combinations of measurement basis choices ($a$, $b$). If the properties of the system were described by a local hidden variable model[2], one would find $|S| \le 2$, whereas any value larger than two indicates a violation of Bell's inequality. The maximum value of $|S|$ allowed by quantum physics is \(2\sqrt{2}\).

In the following, we discuss how we fulfil the requirements outlined here for realizing a Bell test with superconducting circuits closing the locality, fair-sampling and memory loopholes and supporting measurement independence all at the same time.

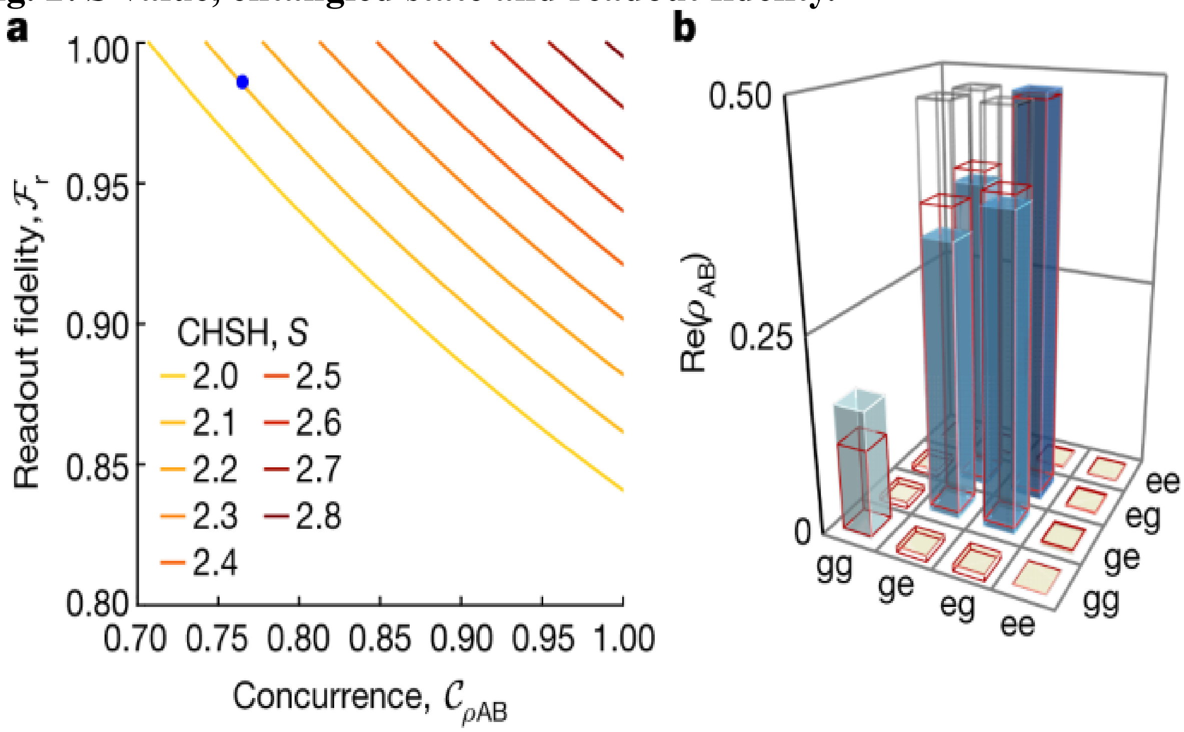
# Requirements

In a Bell test using an entangled pair of qubits, the degree to which Bell's inequality can be violated depends on the concurrence ${\mathcal{C}}$ of the entangled state and the individual qubit-readout fidelity ${{\mathcal{F}}}_{{\rm{r}}}^{({\rm{A}},{\rm{B}})}$. Together, these quantities constrain the maximally achievable Bell parameter to[42]

$${S}^{\max }=2\sqrt{2}{{\mathcal{F}}}_{{\rm{r}}}^{2}{\mathcal{C}}({\rho }_{{\rm{AB}}}).$$
(1)

Thus, the CHSH inequality can only be violated if the average readout fidelity ${{\mathcal{F}}}_{{\rm{r}}}=\sqrt{{{\mathcal{F}}}_{{\rm{r}}}^{{\rm{A}}}{{\mathcal{F}}}_{{\rm{r}}}^{{\rm{B}}}}$ exceeds roughly 84% and the concurrence ${\mathcal{C}}$ exceeds roughly 0.7, so that ${S}^{\max } > 2$, as shown in the contour plot in Fig. 2a.

**Fig. 2: *S* value, entangled state and readout fidelity.**



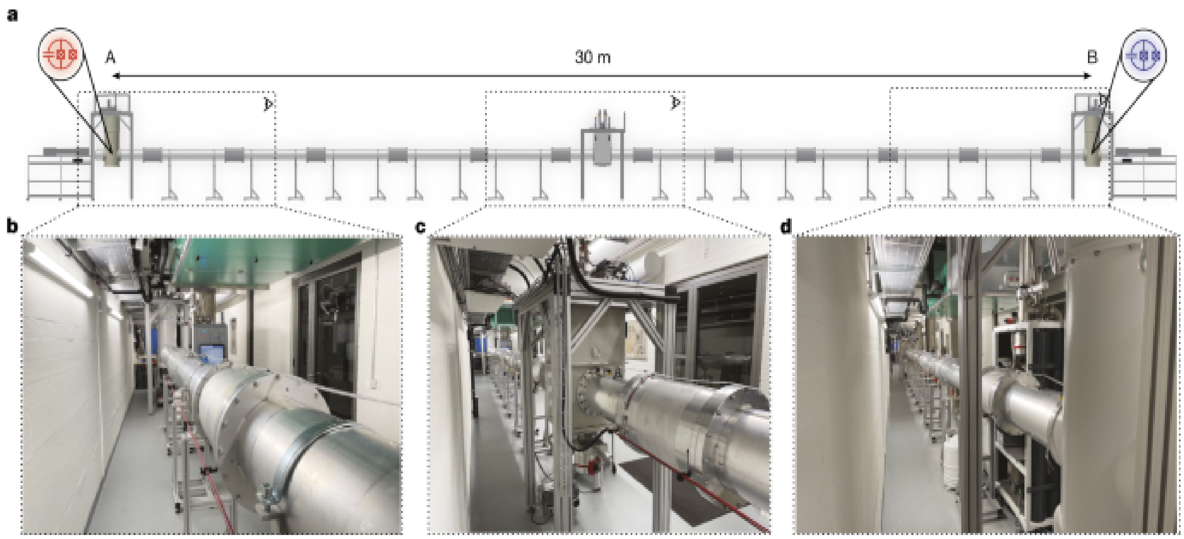

**a**, Calculated *S* value for a Bell test performed in the *xy* basis of the Bloch sphere versus (readout-corrected) Bell state concurrence ${\mathcal{C}}({\rho }_{{\rm{AB}}})$ and average qubit-readout fidelity \

(${{\mathcal{F}}}_{{\rm{r}}}=\sqrt{{{\mathcal{F}}}_{{\rm{r}}}^{{\rm{A}}}{{\mathcal{F}}}_{{\rm{r}}}^{{\rm{B}}}}$). The blue data point indicates the experimentally achieved readout fidelity and concurrence (with correction for readout errors). **b**, Real part of the density matrix $\rho$ of the Bell state $\left|{\psi }^{+}\right\rangle$ reconstructed using quantum state tomography corrected for readout errors. The blue bars indicate the measured, the grey wireframes the ideal values and the red wireframes the results of a master equation simulation.

Addressing the above requirement, previous experiments achieved remote entanglement of superconducting qubits with sufficiently large concurrence in a single dilution refrigerator[12,37,43,44] and in two refrigerators connected across a distance of 5 m using a cryogenic link[14]. In the experiments we present here, we create entanglement over much larger linear distances. In addition, single-shot readout of superconducting qubits was demonstrated in an integration time of 50 ns with fidelity ${{\mathcal{F}}}_{{\rm{r}}}$ exceeding 98% (ref. [13]). In a Bell test that closes the locality loophole, minimizing the duration of the readout reduces the distance $d$ required between the two parties to provide space-like separation.

Given the expected readout time of roughly 50 ns, the time required for choosing the measurement bases at random (roughly 29 ns, Supplementary Information section II), and accounting for a margin for signal propagation times, we chose to build a cryogenic system housing superconducting circuits at a linear physical distance of roughly $d = 30$ m (Fig. 3). This provides a time budget $t_d$ in excess of 100 ns for the Bell test.

**Fig. 3: Cryogenic microwave quantum link.**




**a**, Computer-aided design (CAD) model. **b**–**d**, Photographs of the 30-m-long cryogenic set-up. Dilution refrigerators at each end host the quantum devices that are connected through a waveguide cooled to below 50 mK over the full distance. A central pulse tube cooler provides additional cooling power to the two outermost radiation shields. The photographs are taken at the position of the corresponding eye pictograms shown in **a**. A, (**b**), centre (**c**) and B (**d**).

In our experiments, a pair of dilution refrigerators, one at site A and one at B, each house a superconducting qubit with circuitry for local readout and remote entanglement[12,13,14] cooled down to about 15 mK. In a unique set-up (Fig. 3), we connect the two circuits to each other over a distance of 30 m using a cryogenic quantum microwave channel[14] realized as a superconducting aluminium waveguide. We cool the waveguide to temperatures of a few tens of millikelvin at which its loss is less than 1 dB per km (refs. [14,45]) and its thermal occupation is negligible.

To successfully operate this system, we minimized the heat load at each temperature stage using high-reflectance materials combined with superinsulation for radiation shielding. We designed vertical support structures between the individual shielding stages to minimize thermal conductivity while providing mechanical stability. The system withstands thermal contractions by leveraging flexible thermal connections formed by braids and mobile mechanical supports. We maximize heat flow along the link modules by using high-conductivity copper and minimizing thermal

contact resistances between adjacent link elements. At the midpoint between sites A and B, a pulse tube cooler provides an additional heat sink for the thermal radiation incident on the 50 and 4 K radiation shields. At 30 m in length and with a total mass exceeding 1.3 tons of radiation shields cooled to below 80 K, roughly 90 kg of which are cooled to below 50 mK, this constitutes a large-scale cryogenic system operating at millikelvin temperatures[46]; see Supplementary Information section III for details.

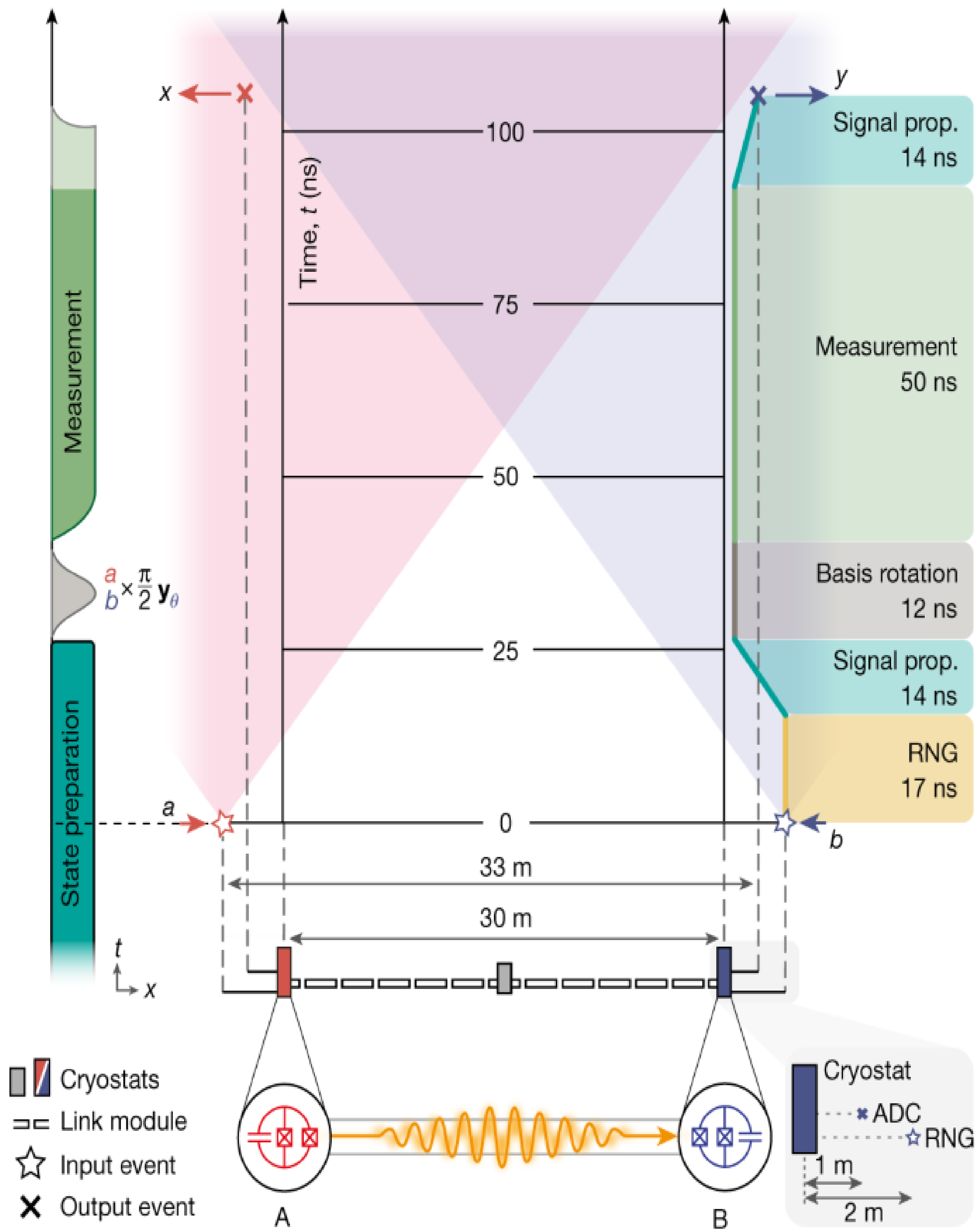
At each of the sites A and B, we operate a transmon-style qubit whose state and transition frequency is controlled on nanosecond time scales using amplitude and phase-modulated microwave pulses and magnetic-flux bias pulses. We read out the state of each qubit using a resonator combined with a Purcell filter[12,13,14]. For the entanglement protocol, we make use of a photon-transfer resonator, also combined with a Purcell filter, which we couple using a coaxial line to the aluminium waveguide connecting the two sites[12,14]. Both qubits and their support circuitry are fabricated on two nominally identical chips (Supplementary Information section IV).

# Bell test

In each individual trial of the Bell test experiment, we deterministically generate a Bell state $\left|{\psi }^{+}\right\rangle =\frac{1}{\sqrt{2}}(\left|ge\right\rangle +\left|eg\right\rangle )$ between the stationary transmon qubits at sites A and B using direct photon exchange[12,47] (Methods). Performing quantum state tomography of the states of the qubits at sites A and B, we experimentally achieve a Bell state fidelity of ${{\mathcal{F}}}_{{\rm{s}}}^{\left|\,\,{\psi }^{+}\right\rangle }=80.4 \%$, corresponding to a concurrence[38] of 0.765 (Fig. 2b) when correcting for readout errors. The experiments were performed with two independent phase synchronized set-ups separated by 30 m (Supplementary Information section V). Without readout error correction we find ${{\mathcal{F}}}_{{\rm{s}}}^{\left|\,\,{\psi }^{+}\right\rangle }=78.9 \%$ and ${\mathcal{C}}=0.689$. The concurrence of the entangled state is sufficiently high for violating Bell's inequality (Fig. 2a) and is on par with previous experiments using the same approach for entangling qubits in a single cryostat[12] and in two cryostats connected across a distance of 5 m

(ref. [14]). The infidelity is dominated by photon loss induced by a circulator used to extract photons from the waveguide for characterization of the entanglement protocol in all three experiments[12,14] (Supplementary Information section VI). As outlined in the Methods section, we rotate each qubit state along the *y* axis of the Bloch sphere to maximize the Bell violation. After entangling the distant qubits at sites A and B, we are set to perform the timing-critical part of the loophole-free Bell test as indicated in the space–time diagram in Fig. 4.

**Fig. 4: Space–time diagram of the experiment.**




The left, vertical time axis schematically shows the microwave pulses applied to the qubits locally at each node. The right axis indicates the duration of the individual Bell test protocol segments: RNG, signal propagation (prop.), qubit basis rotation and measurement. The space–time

location of the start and stop events of a Bell test trial are marked with stars and crosses, respectively. The red and blue regions indicate the future light cones of the start events. The inset on the bottom right indicates the approximate spatial location of the start and stop events in the RNG and ADC, relative to the vertical centre axis of each cryostat.

To generate the input bits a and b for the measurement basis choice we use a RNG at each node[48]. We consider the start event of each trial of a Bell test as being marked in space and time by the earlier of the two events corresponding to the creation of a random number in each RNG. At node A (B), a random number is generated at the location indicated by a red (blue) star in the space–time diagram in Fig. 4 at a distance of about 2 m from the corresponding qubit housed in its dilution refrigerator. The random number *a* (*b*) becomes available as a voltage pulse at the output of the RNG 17.10 ± 0.14 ns after this event (yellow section in Fig. 4). This pulse controls a microwave switch that conditionally passes a microwave basis-rotation pulse to the qubit at A (B). We provide additional information on the random basis choice in the Methods section and in Supplementary Information section II.

To achieve a signal propagation delay of the basis choice pulses applied to each qubit of only 14 ns (first turquoise section in Fig. 4), we pass microwave signals roughly along the line of sight connecting the qubits at A and B from the room-temperature switch through a side-access port into the cryogenic system (Supplementary Information section III). The random basis selection pulse (grey) applied to the qubit has a duration of 12 ns.

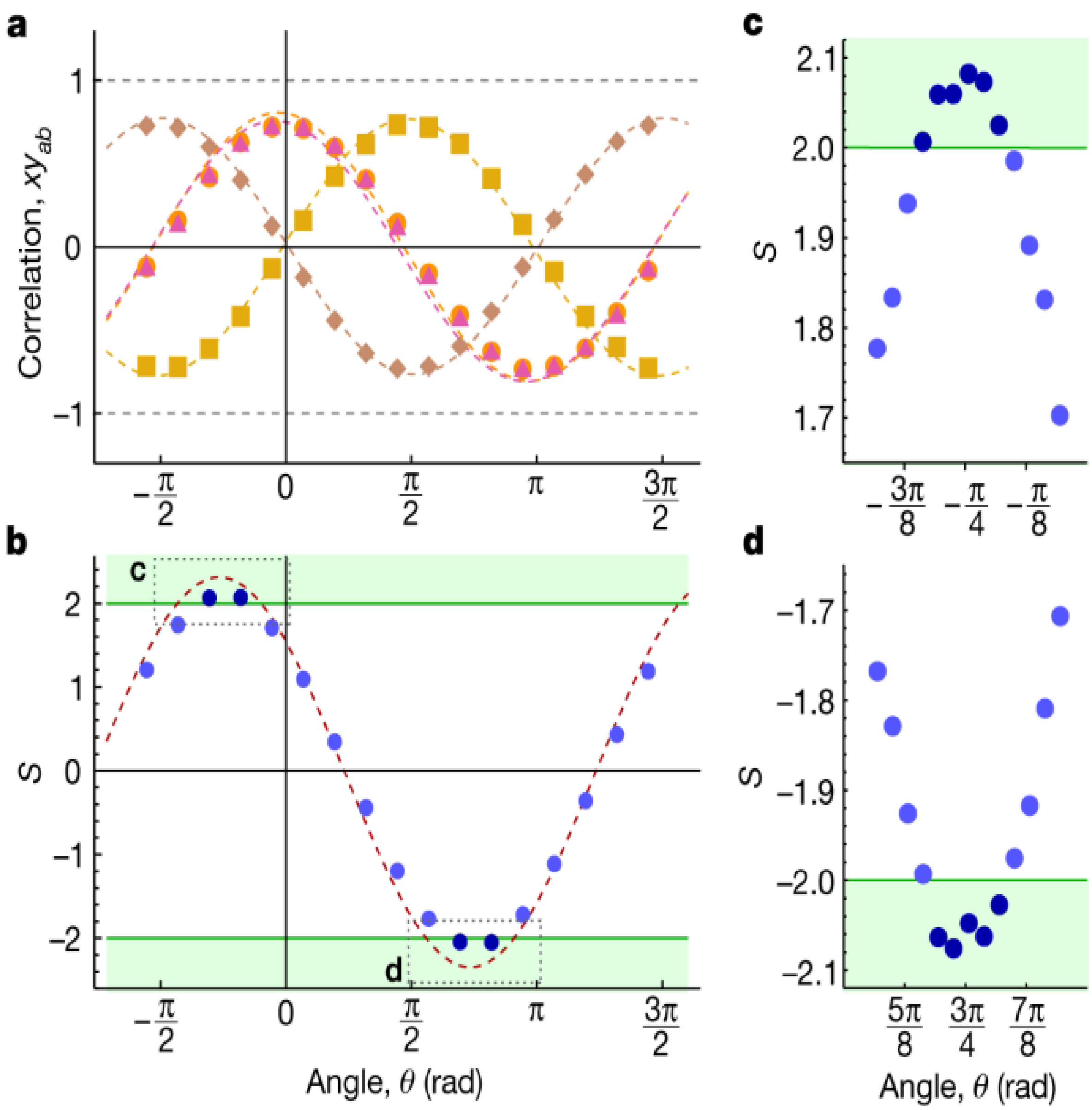
After the microwave pulse has fully rotated both qubit states into the randomly chosen basis, we read out the qubits at A and B by applying a microwave tone to their dedicated readout resonators. We detect the amplitude and phase of the readout pulse after several stages of amplification (green section in Fig. 4), record it with a digitizer (analogue-to-digital converter, ADC) and postprocess the data with a field programmable gate array (FPGA)[13]. We achieve single-shot readout fidelities of $\mathcal{F}_{\rm{r}}^{\rm{A}}=99.05\%$ and $\mathcal{F}_{\rm{r}}^{\rm{B}}=97.60\%$ in only 50 ns integration time (Supplementary Information section VII).

As done with the random basis choice signals, we route the readout signals through the side ports of the dilution refrigerators at sites A and B. In this way, we minimize to 14 ns the propagation delay towards the ADC and FPGA, located at about 1 m physical distance from the qubits, indicated by a cross in Fig. 4. We consider the measurement result to be fixed at time $t_{\times}$, the moment when the last part of the measurement signal arrives at the input of the ADC to be digitized, see Supplementary Information section VIII B for a discussion of this choice.

In each Bell test experiment, we recorded the basis choices $(a, b)$ and the corresponding readout result $(x, y)$ for all $n$ trials. On the basis of these values, we calculated the averages $\langle xy\rangle_{(a,b)}$ for all four combinations of measurement basis choices taking into account all $n$ trials, thus closing the fair-sampling loophole[40,41].

In each of four consecutive experiments, we ran ${n}_{\max }={2}^{20}=1,\,048,\,576$ individual trials of a Bell test for a total time of about 20 minutes. In the first experiment, we swept the angle $\theta$ between the two randomly chosen measurement bases (Fig. 1) across a full period. Plotting $\langle xy\rangle_{(a,b)}$ for all four input bit combinations versus $\theta$ we observe the expected sinusoidal oscillations[4], offset from each other by $\pi/2$ (Fig. 5a). Ideally, $\langle xy\rangle_{(a,b)}$ oscillates between +1 and −1. The observed reduction in contrast is due to the finite concurrence of the initial entangled state and readout errors. We note that we have calibrated out an experimental phase offset of $\theta_0 = 160.0°$ between the two sites (Supplementary Information section IX). We find good agreement between the experimental data and a master equation simulation (Fig. 5a,b and Supplementary Information section VI).

**Fig. 5: Bell inequality violation versus measurement basis offset angle.**




**a**, Quantum correlations $\langle xy \rangle_{(a,b)}$ of individual Bell tests versus offset angle $\theta$. The 17 data points are results of individual Bell tests with \({n}_{max}/17=\mathrm{61,680}\) trials each, incremented by $\theta = \pi/8$. The dashed curves are calculated using a master equation simulation. **b**, Corresponding $S$ values calculated from the data shown in **a**. Points are experimental data, and the dashed red line is extracted from a master equation simulation. Error bars are roughly on the order of the marker size, see text for details. **c**, Measured $S$ values of 13 individual Bell tests, with \({n}_{max}\,/13=\mathrm{80,659}\) trials each, and offset angles around the expected optimum value \({\theta }_{{S}_{\max }}\) incremented by $\theta$

= π/32. **d**, Same as in **c** but for the expected optimum value $\theta_{S_{\min}}$. The green lines in **b**–**d** mark the threshold value $|S| = 2$, and all points in the green shaded region correspond to Bell tests that violate the CHSH inequality.

On the basis of these data, we then calculate the $S$ value as a function of $\theta$ and observe its sinusoidal oscillation with a maximum and a minimum of $S$ occurring at $\theta_{S_{max}}=-\,\pi/4$ and $\theta_{S_{min}}=\pi-\pi/4$ (Fig. 5b), offset by π as expected. Here, we evaluate roughly 60,000 trials for each angle. We find that $|S|$ exceeds two at both values $\theta_{S_{\max/\min}}$, violating the Bell inequality.

Near $\theta_{S_{\max/\min}}$ we perform a set of measurements with step size $\theta = \pi/32$ and determine the $S$ value from roughly 80,000 trials at each value of $\theta$. We observe that several data sets clearly violate the Bell inequality for angles around $\theta_{S_{\max/\min}}$ (Fig. 5c,d). From the data set taken at the offset angle $\theta_{S_{max}}$ we find a maximum violation of $S = 2.082 \pm 0.012 > 2$.

In a final experiment performed at the single angle of $\theta_{S_{max}}$, we acquire Bell test data for $n_{\max}$ trials, yielding $S = 2.0747 \pm 0.0033$, which exceeds two by more than 22 standard deviations. In this experiment with superconducting circuits, we reject the null hypothesis corresponding to Bell's inequality being satisfied with a $P$ value smaller than $10^{-108}$ (Methods), which is small in comparison to $P$ values reported for Bell tests closing all major loopholes in the literature (Supplementary Information section VIII A). The statistical method used here is robust to memory effects (Supplementary Information section X).

Finally, we verify that the locality loophole is closed by measuring the physical distance $d$ separating the two pairs of points in space marked by stars (Fig. 4) defining the start of the Bell test trial at $t = t_{\star} = 0$ from the points in space marked by stars defining the end of the trial at time $t_{\times}$. Using the methods described in the Supplementary Information section XI, we find the shorter of these two distances to be $d = 32.824$ m ± 4.6 mm yielding a time budget of $t_d = 109.489 \pm 0.015$ ns for the Bell test to close the locality

loophole. Using independent measurements, we determine the total duration of the Bell test trial to be $t_\times - t_\star = 107.40 \pm 0.26$ ns $< t_d$ (Supplementary Information section XI), therefore closing the locality loophole with a margin of roughly eight standard deviations. These timing margins are similar to those achieved in loophole-free Bell tests reported in the literature (Supplementary Information section VIII B).

To formulate our conclusion, we assume that we can precisely determine the space–time description of the events at hand and that the RNG produced independent, free random bits. Ultimately, such assumptions constraining the conclusion cannot be fully avoided, even in principle[18]. Under these assumptions, we find that our observation of the violation of Bell's inequality with superconducting circuits is incompatible with an explanation satisfying the principle of local causality.

# Performance and outlook

Previous loophole-free Bell tests using polarization-encoded optical photons as qubits, typically violated Bell's inequality with a lower margin[7,8,9] than our experiment ($S = 2.0747$) whereas atomic and solid-state systems[6,10] realized higher violations. By reducing the loss in the channel connecting the two qubits in our set-up and thus increasing the Bell state fidelity, we expect Bell violations with $S > 2.4$ to be achievable in future experiments while closing all major loopholes. We plan to reach this target by omitting the circulator from the waveguide, as in ref. [37], and using both a low-loss printed circuit board and superconducting microwave cables connecting the sample mount to the waveguide. With these measures, we estimate to reduce the photon loss by up to a factor of four to about 5%. Alternatively, a heralded entangling method, avoiding loss but effectively reducing the repetition rate, can be implemented for the same purpose[12,49]. Such improvements may enable protocols requiring larger Bell violations, such as device-independent quantum key distribution[50], to be executed between superconducting quantum processors connected in a network.

Because the experiment presented here operates at a repetition rate of 12.5 kHz, which is larger than those of loophole-free Bell tests with atomic and

other solid-state systems[6,10], we achieve highly statistically significant Bell inequality violations in only a few minutes. This is similar to experiments performed with polarization-encoded optical photons that have even higher repetition rates[7,8,9]. Further improvements in the repetition rate of our experiment seem feasible, up to the inverse of the duration of the pulse sequence used. In Supplementary Information section VIII we compare in detail the performance metrics of published Bell tests that also used a minimal set of assumptions.

To implement device-independent quantum information processing protocols, it is desirable to simultaneously achieve high Bell violations and high repetition rates. The set-up demonstrated in our experiments provides an interesting combination of those metrics allowing us to visualize implementing a variety of device-independent quantum information processing protocols[26,27,28,29,30] with superconducting circuits, a promising candidate for large-scale quantum computing systems[11,31].

In addition, our experiment demonstrates that quantum information can be transmitted between superconducting circuits housed in cryogenic systems separated by tens of metres, going beyond our previous work on a metre-scale system[14]. Interconnected cryogenic systems may indicate a pathway towards realizing larger scale quantum computing systems using quantum microwave local area networks[51], for example, within a quantum computing centre. The set-up also enables the exploration of non-local quantum physics with degrees of freedom that couple to microwave photons such as mechanical resonators or spins.

# Methods

## Superconducting qubits

For completeness, we discuss a few key features of superconducting quantum electronic circuits here. Superconducting qubits are anharmonic quantum oscillators with circuit parameters chosen to realize resonance frequencies in the gigahertz frequency range (ref. [52] and references therein). The non-linearity of the circuit is provided by an ideally lossless Josephson

element[53], which, in our experiments, is realized as a pair of Josephson tunnel junctions arranged in a superconducting quantum interference device loop shunted by a large capacitor jointly forming a transmon qubit[54]. The effective Hamiltonian of the circuit is governed by a cosinusoidal potential that hosts a set of bound states, the lowest two of that form the computational basis states of the qubit (labelled $\left|g\right\rangle$ and $\left|e\right\rangle$). The second excited state (labelled $| \,f\rangle$) can be used, for example, as an auxiliary state to realize two-qubit gates[55], or, as in this paper, for emitting photons with a controlled temporal mode profile[12,47] when coupling the qutrit strongly to a superconducting resonator (for example, ref. [56] and references therein). Here, as well as elsewhere in this publication, we refer to the quantum bit as a qutrit when we refer to its lowest three energy eigenstates. To minimize thermal excitation of the qubit and also minimize losses in the superconducting materials used to realize the qubits and the 30-m-long waveguide, we operate the devices at temperatures around 15 mK (ref. [25]).

## Generation of the remote, two-qubit entangled state

We generate the Bell state $\left|{\psi }^{+}\right\rangle =(\left|ge\right\rangle +\left|eg\right\rangle )/\sqrt{2}$ between the two remote qubits A and B using a deterministic scheme based on the exchange of a single photon[57], as demonstrated with superconducting circuits, for example, in ref. [12]. In this protocol, qubit A is first entangled with a propagating, ideally time-reversal-symmetric photon in a driven coherent emission process. The propagating photon is then deterministically absorbed in a time-reversed process at qubit B creating the desired entangled state.

The pulse sequence used in the process is about 400 ns long and ends at roughly $t = 16$ ns after the process initiating the random basis choice starts, see Supplementary Information section VI for details. We create the entangled state as an input resource to the presented Bell test experiment. We consider the entanglement process itself to be independent of the timing constraints of the loophole-free Bell test (Methods and Supplementary Information sections II and XI).

We characterize the created Bell state using quantum state tomography (Fig. 2b) for which we rotate the density matrix in postprocessing by an angle $\theta_0$ around the $z$ axis to maximize the Bell state fidelity. $\theta_0$ is the experimental offset angle between the two set-ups A and B (Supplementary Information section IX). We also perform a master equation simulation of the Bell state generation protocol and find good agreement with the experimental data (red wireframes in Fig. 2b), characterized by the small trace distance $\sqrt{{\rm{Tr}}(| \rho -{\rho }_{{\rm{sim}}}{| }^{2}\,)}=0.077$.

## Optimizing the measurement basis

Photon loss and qubit decay are the dominant mechanisms reducing the fidelity of the experimentally created Bell state. These processes create an asymmetry between the qubit excited and ground states participating in the Bell test. To reduce this detrimental effect and maximize the $S$ value, we perform the Bell test by choosing measurement bases in the $xy$ plane of the Bloch sphere, where each basis is affected equally by photon loss and qubit decay. To do so, we apply a $(\pi/2)_x$ basis-rotation pulse to qubit A, and a $(\pi/2)_{x+\theta}$ pulse to qubit B in each experimental trial just after the $\left|{\psi }^{+}\right\rangle$ Bell state preparation and before applying the pulse implementing the random measurement basis choice. Here, $\theta$ denotes the angle between the two randomly chosen measurement bases, as introduced in the main text. Effectively, this pulse sequence generates the Bell state $\left|{\varphi }^{+}\right\rangle =(\left|gg\right\rangle +\left|ee\right\rangle )/\sqrt{2}$.

## Random basis selection and measurement independence

As in previous loophole-free Bell tests[6,7,8,10], we use well-characterized, fast physical RNGs[48] to support the measurement independence assumption. The basis choice bits a and b are generated by a dedicated RNG at each node. Each RNG contains eight quantum entropy sources, each composed of a laser to produce phase-randomized pulses, an interferometer, fast linear detection, one-bit digitization and a parity calculator. We assume that the extracted random bits are independent from all previous events.

Supplementary Information section II describes support for this assumption and in Supplementary Information section X we perform a statistical analysis.

The RNG output bit is encoded as a voltage, and controls a single-pole single-throw microwave switch, which (with input high) blocks or (with input low) passes a microwave $(\pi/2)_y$ pulse to induce a basis change of the corresponding qubit. We discuss the basis selection further in Supplementary Information section II.

Because every choice of measurement leads to a recorded experimental trial in our experiment, as in refs. [7,8], the RNG properties offer direct support for the validity of the measurement independence condition. This differs from the situation in event-ready Bell tests[6,10], in which a heralding event—the result of a joint photon measurement at a middle station—determines whether or not a given experimental trial is recorded for analysis. Such selection opens the possibility for the heralding event to select trials as a function of the measurement choices, if these are not space-like separated from the joint measurement. In this situation, the measurement independence condition can be violated even in presence of perfect RNGs. For this reason, in event-ready Bell tests it is important that the heralding event is space-like separated from the events marking the creation of the random-input bits. Because we use a deterministic entanglement generation protocol that is not dependent on the outcome of any measurement, such considerations related to a heralding event do not play a role in our experiment.

Note that when focusing on specific families of local hidden variable models, measurement independence can sometimes be supported by space-like separation alone. This is the case for the model introduced in Scheidl et al.[58], and relevant for Bell tests with entangled-photon pairs, where the hidden variable $\lambda$ is assumed to be created at the photon pair generation event independently of any past event. With that assumption, space-like separation between the pair generation and the measurement choice events guarantees measurement independence. This space-like separation is achieved also in refs. [7,8]. Assuming $\lambda$ is produced along with the photon pairs, but relaxing the assumption that $\lambda$ is independent of past events, a

photonic Bell test with the same space-like separation between the pair generation and the settings choice events can exclude local causal models in which the photon pairs influence the measurement choices, but not local causal models in which earlier events influence both the photons and the basis choices. Because such past influences could be arbitrarily far in the past, space–time conditions cannot be used to fully support the assumption of measurement independence beyond the model of Scheidl et al.[58], even in photonic Bell tests. For this reason, we do not attempt to create a space-like separation between the entanglement generation and the basis choices, but rather support the assumption of measurement independence by using well-characterized RNGs[48], as has been done in previous Bell tests since the pioneering work of Weihs et al.[20].

### The *P* value as a statistical metric

Early Bell test experiments typically used the standard deviation as a metric to discuss the statistical significance of an observed Bell inequality violation. This approach, however, comes with two limitations. The first is that when using the notion of standard deviation, we implicitly assume that the underlying measurement data are Gaussian distributed. This assumption is only justified in the limit of infinitely many trials, but in experiments a finite number of trials are executed. A statistical analysis of the Bell test based on standard deviations may therefore overestimate the statistical significance of the result[5,59]. The second limitation is that the notion of standard deviation relies on the assumption that the result of the $k$th trial is independent of the basis choices and measurement results of the previous $k - 1$ trials, which opens up the memory loophole[18]. These two limitations can be addressed by the statistical analysis of the result through the calculation of a *P* value according to a method that does not rely on any of the aforementioned assumptions. Therefore, the calculation of *P* values in the context of Bell tests is now an established practice[6,7,8,9,10]. In this context, the *P* value is a metric of the probability with which data as extreme as those observed could have been produced by a local causal model (see Supplementary Information section X for details).

## Data availability

All data are available from the corresponding authors upon reasonable request.

# References


1. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **48**, 696 (1935).

2. Bell, J. S. On the Einstein Podolsky Rosen paradox. *Physics* **1**, 195 (1964).

3. Bell, J. S. in *Speakable and Unspeakable in Quantum Mechanics: Collected Papers on Quantum Philosophy* 2nd edn, Ch. 24, 232–248 (Cambridge Univ. Press, 2004).

4. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880 (1969).

5. Brunner, N., Cavalcanti, D., Pironio, S., Scarani, V. & Wehner, S. Bell nonlocality. *Rev. Mod. Phys.* **86**, 419 (2014).

6. Hensen, B. et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).

7. Giustina, M. et al. Significant-loophole-free test of Bell's theorem with entangled photons. *Phys. Rev. Lett.* **115**, 250401 (2015).

8. Shalm, L. K. et al. Strong loophole-free test of local realism. *Phys. Rev. Lett.* **115**, 250402 (2015).

9. Li, M.-H. et al. Test of local realism into the past without detection and locality loopholes. *Phys. Rev. Lett.* **121**, 080404 (2018).

10. Rosenfeld, W. et al. Event-ready Bell test using entangled atoms simultaneously closing detection and locality loopholes. *Phys. Rev. Lett.* **119**, 010402 (2017).

11. Arute, F. et al. Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).

12. Kurpiers, P. et al. Deterministic quantum state transfer and remote entanglement using microwave photons. *Nature* **558**, 264–267 (2018).

13. Walter, T. et al. Rapid, high-fidelity, single-shot dispersive readout of superconducting qubits. *Phys. Rev. Appl.* **7**, 054020 (2017).

14. Magnard, P. et al. Microwave quantum link between superconducting circuits housed in spatially separated cryogenic systems. *Phys. Rev. Lett.* **125**, 260502 (2020).

15. Scarani, V. *Bell Nonlocality* (Oxford Univ. Press, 2019).

16. Freedman, S. J. & Clauser, J. F. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* **28**, 938 (1972).

17. Aspect, A., Grangier, P. & Roger, G. Experimental realization of Einstein-Podolsky-Rosen-Bohm gedankenexperiment: a new violation of Bell's inequalities. *Phys. Rev. Lett.* **49**, 91 (1982).

18. Larsson, J.-A. Loopholes in Bell inequality tests of local realism. *J. Phys. A: Math. Theor.* **47**, 424003 (2014).

19. Aspect, A., Dalibard, J. & Roger, G. Experimental test of Bell's inequalities using time- varying analyzers. *Phys. Rev. Lett.* **49**, 1804 (1982).

20. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. & Zeilinger, A. Violation of Bell's inequality under strict Einstein locality conditions. *Phys. Rev. Lett.* **81**, 5039 (1998).

21. Rowe, M. A. et al. Experimental violation of a Bell's inequality with efficient detection. *Nature* **409**, 791–794 (2001).

22. Aspect, A. Closing the door on Einstein and Bohr's quantum debate. *Physics* **8**, 123 (2015).

23. Acín, A. et al. Device-independent security of quantum cryptography against collective attacks. *Phys. Rev. Lett.* **98**, 230501 (2007).

24. Šupić, I. & Bowles, J. Self-testing of quantum systems: a review. *Quantum* **4**, 337 (2020).

25. Sekatski, P., Bancal, J.-D., Wagner, S. & Sangouard, N. Certifying the building blocks of quantum computers from Bell's theorem. *Phys. Rev. Lett.* **121**, 180505 (2018).

26. Vazirani, U. & Vidick, T. Fully device-independent quantum key distribution. *Phys. Rev. Lett.* **113**, 140501 (2014).

27. Colbeck, R. *Quantum and Relativistic Protocols for Secure Multi-Party Computation*. PhD thesis, Univ. Cambridge (2009).

28. Pironio, S. et al. Random numbers certified by Bell's theorem. *Nature* **464**, 1021–1024 (2010).

29. Colbeck, R. & Renner, R. Free randomness can be amplified. *Nat. Phys.* **8**, 450–453 (2012).

30. Kessler, M. & Arnon-Friedman, R. Device-independent randomness amplification and privatization. *IEEE J. Selected Areas Inf. Theory* **1**, 568–584 (2020).

31. Krinner, S. et al. Realizing repeated quantum error correction in a distance-three surface code. *Nature* **605**, 669–674 (2022).

32. Leggett, A. J. Macroscopic quantum systems and the quantum theory of measurement. *Progr. Theor. Phys. Suppl.* **69**, 80–100 (1980).

33. Clarke, J. & Wilhelm, F. K. Superconducting quantum bits. *Nature* **453**, 1031–1042 (2008).

34. Martinis, J. M., Devoret, M. H. & Clarke, J. Energy-level quantization in the zero-voltage state of a current-biased Josephson junction. *Phys. Rev. Lett.* **55**, 1543 (1985).

35. Ansmann, M. et al. Violation of Bell's inequality in Josephson phase qubits. *Nature* **461**, 504–506 (2009).

36. The BIG Bell Test Collaboration. Challenging local realism with human choices. *Nature* **557**, 212–216 (2018).

37. Zhong, Y. P. et al. Violating Bell's inequality with remotely connected superconducting qubits. *Nat. Phys.* **15**, 741–744 (2019).

38. Wootters, W. K. Entanglement of formation of an arbitrary state of two qubits. *Phys. Rev. Lett.* **80**, 2245 (1998).

39. Horodecki, R., Horodecki, P. & Horodecki, M. Violating Bell inequality by mixed spin-12 states: necessary and sufficient condition. *Phys. Lett. A* **200**, 340 (1995).

40. Garg, A. & Mermin, N. D. Detector inefficiencies in the Einstein-Podolsky-Rosen experiment. *Phys. Rev. D* **35**, 3831 (1987).

41. Eberhard, P. H. Background level and counter efficiencies required for a loophole-free Einstein-Podolsky-Rosen experiment. *Phys. Rev. A* **47**, R747 (1993).

42. Magnard, P. *Meter-Scale Microwave Quantum Networks for Superconducting Circuits*. PhD thesis, ETH Zurich (2021).

43. Axline, C. et al. On-demand quantum state transfer and entanglement between remote microwave cavity memories. *Nat. Phys.* **14**, 705–710 (2018).

44. Campagne-Ibarcq, P. et al. Deterministic remote entanglement of superconducting circuits through microwave two-photon transitions. *Phys. Rev. Lett.* **120**, 200501 (2018).

45. Kurpiers, P., Walter, T., Magnard, P., Salathe, Y. & Wallraff, A. Characterizing the attenuation of coaxial and rectangular microwave-frequency waveguides at cryogenic temperatures. *EPJ Quant. Technol.* **4**, 8 (2017).

46. D'Addabbo, A. et al. The CUORE cryostat. *J. Low Temp. Phys.* **193**, 867–875 (2018).

47. Pechal, M. et al. Microwave-controlled generation of shaped single photons in circuit quantum electrodynamics. *Phys. Rev. X* **4**, 041010 (2014).

48. Abellán, C., Amaya, W., Mitrani, D., Pruneri, V. & Mitchell, M. W. Generation of fresh and pure random numbers for loophole-free Bell tests. *Phys. Rev. Lett.* **115**, 250403 (2015).

49. Burkhart, L. D. et al. Error-detected state transfer and entanglement in a superconducting quantum network. *PRX Quantum* **2**, 030321 (2021).

50. Barrett, J., Hardy, L. & Kent, A. No signalling and quantum key distribution. *Phys. Rev. Lett.* **95**, 010503 (2005).

51. Peñas, G. F., Puebla, R., Ramos, T., Rabl, P. & García-Ripoll, J. J. Universal deterministic quantum operations in microwave quantum links. *Phys. Rev. Applied* **17**, 054038 (2022).

52. Kjaergaard, M. et al. Superconducting qubits: current state of play. *Ann. Rev. Condens. Matter Phys.* **11**, 369–395 (2020).

53. Josephson, B. D. Possible new effects in superconductive tunnelling. *Phys. Lett.* **1**, 251–253 (1962).

54. Koch, J. et al. Charge-insensitive qubit design derived from the Cooper pair box. *Phys. Rev. A* **76**, 042319 (2007).

55. DiCarlo, L. et al. Demonstration of two-qubit algorithms with a superconducting quantum processor. *Nature* **460**, 240–244 (2009).

56. Blais, A., Grimsmo, A. L., Girvin, S. M. & Wallraff, A. Circuit quantum electrodynamics. *Rev. Mod. Phys.* **93**, 025005 (2021).

57. Cirac, J. I., Zoller, P., Kimble, H. J. & Mabuchi, H. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78**, 3221 (1997).

58. Scheidl, T. et al. Violation of local realism with freedom of choice. *Proc. Natl Acad. Sci. USA* **107**, 19708 (2010).

59. Zhang, Y., Glancy, S. & Knill, E. Asymptotically optimal data analysis for rejecting local realism. *Phys. Rev. A* **84**, 062118 (2011).

# Acknowledgements

We thank M. Frey, N. Kohli, R. Schlatter, A. Fauquex, R. Keller, B. Dönmez, M. Hinderling, S. Wili, F. Marxer, A. Schwarzer and J.-A. Agner for technical support on setting up the laboratory, and designing, building and testing the cryogenic system. We thank J. Herrmann, L. Raabe, N. Mostaan, E. Portolés, M. Ruckriegel and J. Heinsoo for their contributions to software and electronics. We thank A. Aspect, G. Blatter, N. Gisin, R. Hanson, A. Imamoglu, R. Renner and R. Wolf for commenting on an early version of the manuscript. The work at ETH Zurich was funded by the European Research Council through the 'Superconducting Quantum Networks' (SuperQuNet) project, by the European Union's Horizon 2020 FET-Open project SuperQuLAN (grant no. 899354), by the National Centre of Competence in Research 'Quantum Science and Technology' (NCCR QSIT), a research instrument of the Swiss National Science Foundation, and by ETH Zurich. B.R. and A.B. acknowledge support from the Natural Sciences and Engineering Research Council of Canada, the Canada First Research Excellence Fund and from the Vanier Canada Graduate Scholarships. J.-D.B. and N.S. acknowledge support by the Institut de Physique Théorique, Commissariat à l'Energie Atomique et aux Energies Alternatives, by the European High-Performance Computing Joint Undertaking under grant agreement no. 101018180 and project name HPCQS and by a French national quantum initiative managed by Agence Nationale de la Recherche in the framework of France 2030 with the references ANR-22-PETQ-0007, project name EPIQ and ANR-22-PETQ-0009, project name DIQKD. M.W.M. acknowledges support by NextGenerationEU (grant no. PRTR-C17.I1) and by projects SAPONARIA (grant no. PID2021-123813NB-I00) and MARICHAS (grant no. PID2021-126059OA-I00), by 'Severo Ochoa' Center of Excellence CEX2019-000910-S, Generalitat de Catalunya through the CERCA program, AGAUR

grant no. 2021-SGR-01453, by Fundació Privada Cellex and by Fundació Mir-Puig.

# Funding

Open access funding provided by Swiss Federal Institute of Technology Zurich.

# Author information

Author notes

1. Paul Magnard

   Present address: Alice and Bob, Paris, France

2. Philipp Kurpiers

   Present address: Rohde and Schwarz, Munich, Germany

3. Adrian Copetudo

   Present address: Centre for Quantum Technologies, National University of Singapore, Singapore, Singapore

## Authors and Affiliations

1. Department of Physics, ETH Zurich, Zurich, Switzerland

   Simon Storz, Josua Schär, Anatoly Kulikov, Paul Magnard, Philipp Kurpiers, Janis Lütolf, Theo Walter, Adrian Copetudo, Kevin Reuer, Abdulkadir Akin, Jean-Claude Besse, Mihai Gabureac, Graham J. Norris, Andrés Rosario & Andreas Wallraff

2. Quside Technologies S.L., Castelldefels, Spain

   Ferran Martin, José Martinez, Waldimar Amaya & Carlos Abellan

3. ICFO - Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, Castelldefels (Barcelona), Spain

   Morgan W. Mitchell

4. ICREA - Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

   Morgan W. Mitchell

5. Institute of Theoretical Physics, University of Paris-Saclay, CEA, CNRS, Gif-sur-Yvette, France

   Jean-Daniel Bancal & Nicolas Sangouard

6. Department of Physics, Yale University, New Haven, CT, USA

   Baptiste Royer

7. Institut quantique and Départment de Physique, Université de Sherbrooke, Sherbrooke, Québec, Canada

   Baptiste Royer & Alexandre Blais

8. Canadian Institute for Advanced Research, Toronto, Ontario, Canada

   Alexandre Blais

9. Quantum Center, ETH Zurich, Zurich, Switzerland

   Andreas Wallraff

## Contributions

S.S., J.S., A.K. and P.M. planned and performed the experiment and analysed the data. S.S., P.M., T.W. and J.S. designed and tested the quantum devices. J.S., J.L., S.S., P.M., P.K. and A.W. devised and tested the cryogenic set-up. J.-C.B., M.G., G.J.N., T.W. and A.R. fabricated the devices. P.M. and S.S. integrated the measurement basis choice scheme. A.K., A.C., S.S., J.S.

and P.M. developed and implemented the set-up synchronization scheme and the verification procedure for closing the locality loophole. S.S., A.C., K.R., P.M. and A.A. developed the experiment control and FPGA-based data analysis code. B.R., S.S. and A.B. ran the master equation simulations. F.M., J.M., W.A., M.W.M., C.A. and P.M. developed and tested the RNGs. J.-D.B. and N.S. assisted with the statistical analysis and the manuscript. S.S., J.S. and P.M. provided the figures for the manuscript. S.S. and A.W. wrote the manuscript with major contributions from A.K., J.S., J.-D.B., N.S. and M.W.M. and input from all authors. A.W. supervised the project.

### Corresponding authors

Correspondence to Simon Storz or Andreas Wallraff.

## Ethics declarations

### Competing interests

The authors declare no competing interests.

## Peer review

### Peer review information

*Nature* thanks Peter Bierhorst, Yanbao Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

## Additional information

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Supplementary information

## Supplementary Information

This Supplementary Information file contains 11 sections, Supplementary Figs. 1–12, Tables 1–6 and additional references.

## Peer Review File

# Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

# About this article

## Cite this article

Storz, S., Schär, J., Kulikov, A. *et al.* Loophole-free Bell inequality violation with superconducting circuits. *Nature* **617**, 265–270 (2023). https://doi.org/10.1038/s41586-023-05885-0

- Received: 22 August 2022
- Accepted: 24 February 2023
- Published: 10 May 2023

- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05885-0

### Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

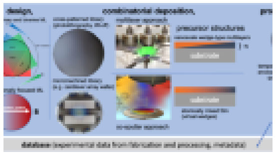
## Superconducting qubits cover new distances

- Marissa Giustina

Nature News & Views 10 May 2023

---

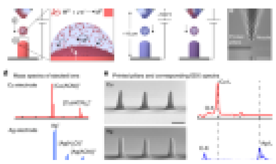
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05885-0



## Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods

10 July 2019

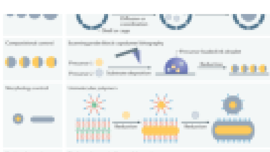
Alfred Ludwig



## Multi-metal electrohydrodynamic redox 3D printing at the submicron scale

23 April 2019

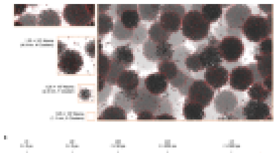
Alain Reiser, Marcus Lindén, … Ralph Spolenak



## Nanoreactors for particle synthesis

13 January 2022

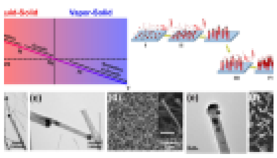
Jordan H. Swisher, Liban Jibril, … Chad A. Mirkin



## [Chemistry-mediated Ostwald ripening in carbon-rich C/O systems at extreme conditions]

17 March 2022

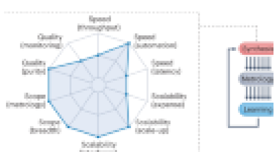
Rebecca K. Lindsey, Nir Goldman, … Sorin Bastea



## [In situ synthesis of hierarchically-assembled three-dimensional ZnS nanostructures and 3D printed visualization]

10 October 2022

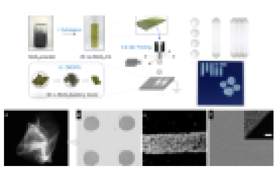
Taehwan Lim, Seung Kwon Seol, … Jung Han Kim



## [Combinatorial synthesis for AI-driven materials discovery]

09 March 2023

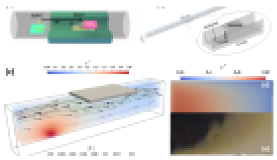
John M. Gregoire, Lan Zhou & Joel A. Haber



[**Additive manufacturing assisted van der Waals integration of 3D/3D hierarchically functional nanostructures**]

15 July 2020

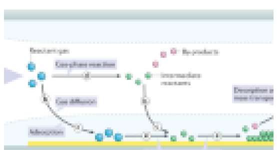
Jui-Han Fu, Ang-Yu Lu, … Vincent Tung



[**Multiscale framework for simulation-guided growth of 2D materials**]

14 September 2018

Kasra Momeni, Yanzhou Ji, … Long-Qing Chen



[**Chemical vapour deposition**]

14 January 2021

Luzhao Sun, Guowen Yuan, … Zhongfan Liu

- Article
- Open Access
- Published: 10 May 2023

# High-throughput printing of combinatorial materials from

# aerosols


- Minxiang Zeng,
- Yipu Du,
- Qiang Jiang,
- Nicholas Kempf,
- Chen Wei,
- Miles V. Bimrose,
- A. N. M. Tanvir,
- Hengrui Xu,
- Jiahao Chen,
- Dylan J. Kirsch,
- Joshua Martin,
- Brian C. Wyatt,
- Tatsunori Hayashi,
- Mortaza Saeidi-Javash,
- Hirotaka Sakaue,
- Babak Anasori,
- Lihua Jin,
- Michael D. McMurtrey &
- …
- Yanliang Zhang

Show authors


*Nature* **volume 617**, pages 292–298 (2023)


- 11 Altmetric

- Metrics details


## Abstract

The development of new materials and their compositional and microstructural optimization are essential in regard to next-generation technologies such as clean energy and environmental sustainability.

However, materials discovery and optimization have been a frustratingly slow process. The Edisonian trial-and-error process is time consuming and resource inefficient, particularly when contrasted with vast materials design spaces[1]. Whereas traditional combinatorial deposition methods can generate material libraries[2,3], these suffer from limited material options and inability to leverage major breakthroughs in nanomaterial synthesis. Here we report a high-throughput combinatorial printing method capable of fabricating materials with compositional gradients at microscale spatial resolution. In situ mixing and printing in the aerosol phase allows instantaneous tuning of the mixing ratio of a broad range of materials on the fly, which is an important feature unobtainable in conventional multimaterials printing using feedstocks in liquid–liquid or solid–solid phases[4,5,6]. We demonstrate a variety of high-throughput printing strategies and applications in combinatorial doping, functional grading and chemical reaction, enabling materials exploration of doped chalcogenides and compositionally graded materials with gradient properties. The ability to combine the top-down design freedom of additive manufacturing with bottom-up control over local material compositions promises the development of compositionally complex materials inaccessible via conventional manufacturing approaches.

# Main

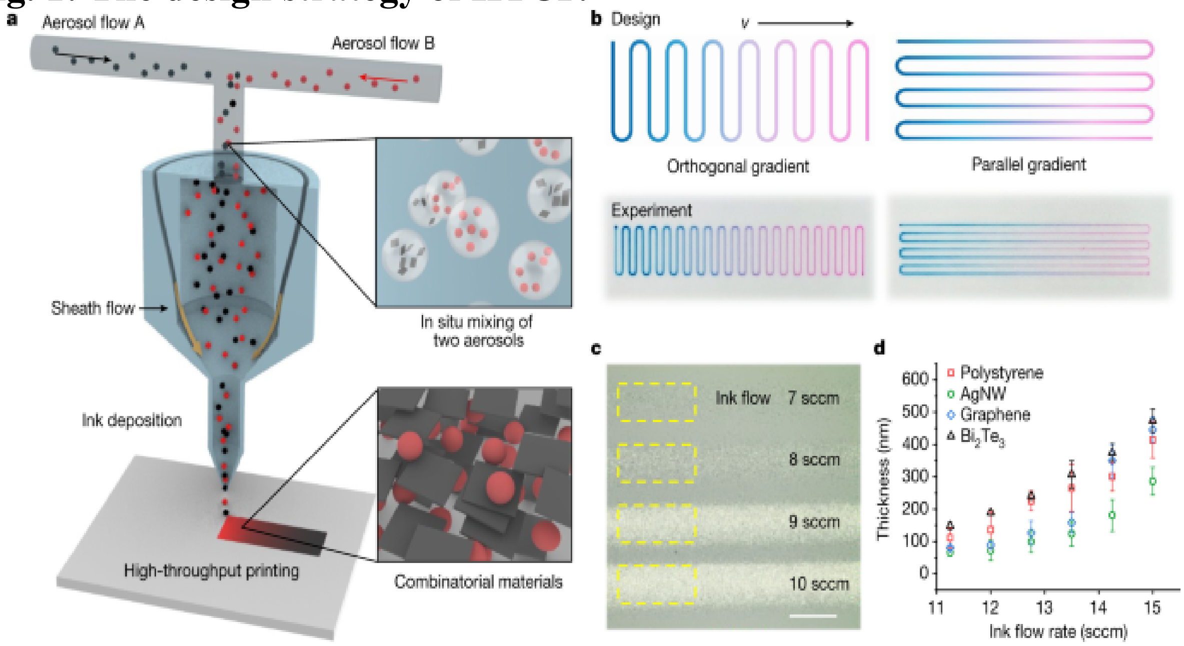
Materials hold pivotal roles in many scientific and technological innovations, and progress in developing new materials is key to the pursuit of solutions to grand societal challenges. Combinatorial material depositions (for example, cosputtering) have enabled rapid screening of new materials for electronics, magnetics, optics and energy-related applications[7]. The sample-rich feature of these combinatorial material libraries facilitates elucidation of the composition–structure–property relationship and enables the rapid screening of materials over a vast range of compositions. Nevertheless, the intrinsic high-energy nature of laser or plasma excludes many materials (for example, colloidal particles, thermosensitive polymers) from use in the development of universal combinatorial material libraries. Additive manufacturing has emerged as a versatile method to fabricate materials of complex structure using micro- and nanoscale building blocks[8,9,10]. Recently several printing approaches, including inkjet printing,

electrochemical printing and electrohydrodynamic redox printing, have been proposed for the fabrication of material libraries[11,12,13]. However, these methods still suffer from limited materials options and challenges in regard to the universal combination of different materials and the production of gradient material libraries, due to the lack of fast mixing mechanisms and the inability to rapidly vary mixing ratios.

For an ideal interdiffusion system, low fluid viscosity and minimal size of diffusion units are desired, which leads us to investigate the potential of using aerosols for in situ mixing and printing. Previous research on multimaterial aerosol jet printing has made steady progress in the development of functional materials and devices[14,15], although aerosol-based printing of combinatorial gradient materials remains challenging. During aerosol-based printing, the material deposition rate can be affected by several parameters (aerosol ink flow rate, sheath gas flow rate, printing speed, atomizing voltage and so on)[16,17], and the interplay of these printing parameters complicates aerosol mixing and deposition during printing. Unoptimized ink formulation and printing conditions may lead to unstable jetting, which can introduce uncertainty in aerosol-based printing. To understand the collective behaviour of aerosol mixing and the combinatorial printing process, we systematically investigated ink formulation, aerosol mixing and interaction, and printing parameter optimization by combining both experimental techniques (for example, fast camera imaging) and computational fluid dynamics (CFD) simulations. To achieve aerosol-based mixing and printing, our high-throughput combinatorial printing (HTCP) approach begins with atomization of two (or multiple) inks into aerosols containing microscale ink droplets, where the combined ink streams are then mixed in a single nozzle and aerodynamically focused by a co-flowing sheath gas before deposition (Fig. 1a). The aerosol jet printhead with nozzles of various sizes is applied, delivering fine features with spatial resolution as low as around 20 μm in the *x*–*y* plane and deposition thickness as low as approximately 100 nm (Supplementary Figs. 1 and 2). To generate a one-dimensional (1D) gradient material library we investigated two printing strategies—orthogonal versus parallel gradient printing (Fig. 1b). Although both approaches can generate gradient films, we found that orthogonal printing tends to be more versatile because it can tolerate a wide range of printing speeds. By contrast, a high printing speed in parallel gradient

mode may lead to an undesired deposition delay that causes inaccurate ink mixing and deposition (Fig. 1b). By continuously varying the ink-mixing ratio through orthogonal printing, the compositional variation of printed materials can be achieved in a fine-gradient manner without the requirement for clean-room facilities (Supplementary Video 1 and Supplementary Table 1).

**Fig. 1: The design strategy of HTCP.**




**a**, Schematic illustration of the combinatorial printing method based on in situ aerosol mixing. **b**, Orthogonal and parallel gradient printing design strategies, and corresponding printed gradient patterns using blue ink (food dye Blue 1) and red ink (rhodamine B), demonstrating a compositional–modulation feature. **c**, Optical microscopy images showing the impact of aerosol ink flow rate on the deposited materials. Scale bar, 100 µm. **d**, Printed material thickness versus flow rate of various inks (polystyrene, AgNW, graphene and $Bi_2Te_3$). Error bars represent s.d. from four experimental replicates. sccm, standard cubic centimetres per minute.

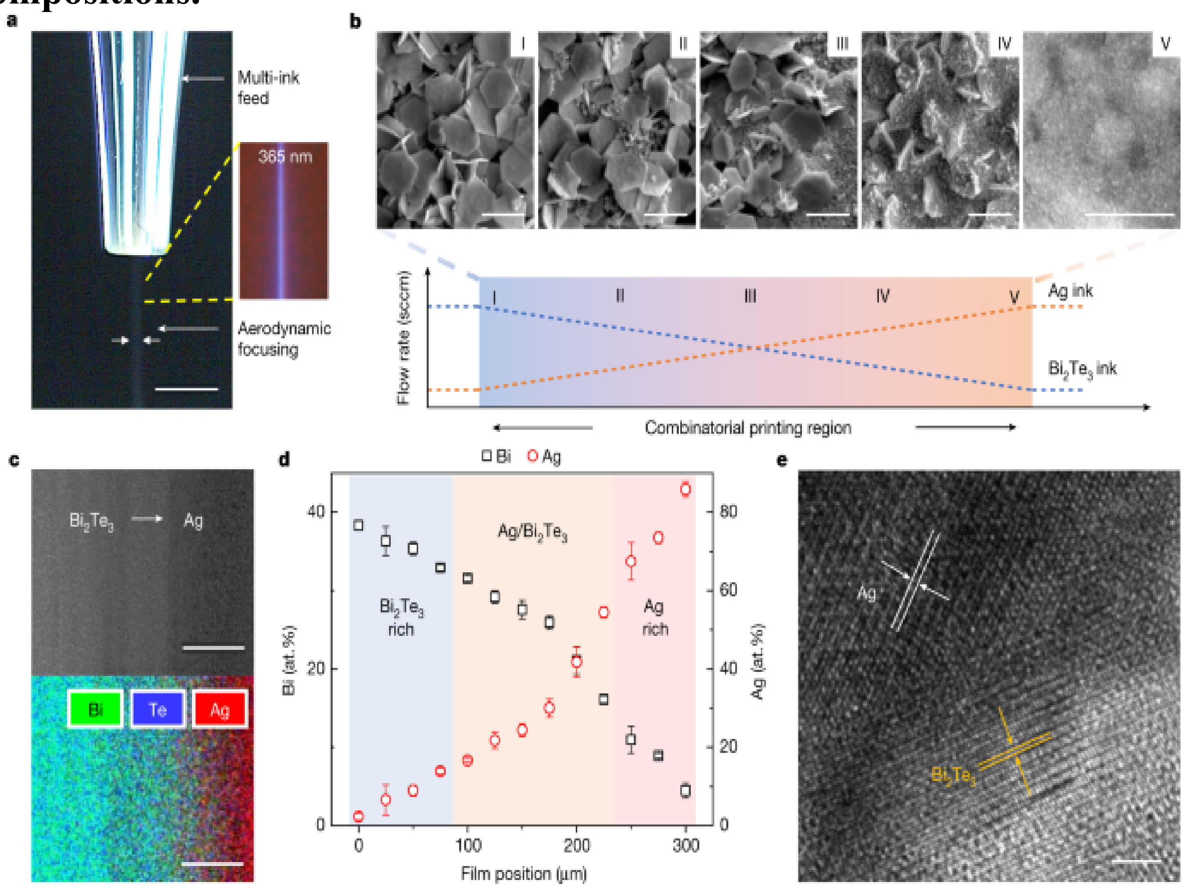
# Combinatorial mixing and printing

The printing of gradient material libraries relies on two hypotheses: (1) the controllable deposition of two materials via individual modulation of two

ink flow rates; and (2) the mixing of two ink aerosols on the fly. We first evaluated the effect of ink flow rates on material deposition. As shown in Fig. 1c, material deposition rate can be controlled by adjusting ink flow rate within the stable jetting range, where the resulting deposition thicknesses of printed films increase with ink flow rate monotonically. In an optimized range of ink flow rates (Fig. 1d), we found that this monotonic trend can be applied to a variety of nanomaterial inks including Ag nanowires (AgNW), graphene, $Bi_2Te_3$ and polystyrene, although an extremely high aerosol flow rate may lead to unstable jetting[17,18]. Other printing parameters have been systematically investigated to optimize printing processes and realize high printing reproducibility and stability. For printing and ink parameters see Supplementary Fig. 3 and Supplementary Tables 2–5; for ink priming, Supplementary Fig. 4; for printing stability, Supplementary Fig. 5; for batch-to-batch reproducibility, Supplementary Fig. 6; for printing uncertainty, Supplementary Fig. 7; for correlation matrix of parameters, Supplementary Fig. 8; for aerosol ink flow range, Supplementary Fig. 9; for ink effect on surface smoothness, Supplementary Fig. 10; for printing reproducibility of gradient samples, Supplementary Figs. 11 and 12; for substrate effect, Supplementary Figs. 13 and 14; and for typical flow, Supplementary Fig. 2a. In particular, we analysed the role of printing process variables to show the relationship between process variability and underlying mechanisms. To minimize process variability due to changes in ink properties, ink formulation (for example, solvents and surfactants) needs to be engineered to ensure long-term chemical stability[19] and colloidal stability[20]. It is also important to optimize printing process parameters to control the key flow characteristics and ensure stable jetting with low variability for aerosol-based printing. In addition, fast camera imaging was utilized to understand the behaviour of aerosol jetting of inks, where we observed the effect of Saffman force on aerosol droplets with a strong collimating effect that helps narrow the passage of ink aerosols (for sheath flow effect see Supplementary Fig. 15 and Supplementary Video 2; for time-dependent jetting, Supplementary Fig. 16; and for aerosols of fluorescent inks, Supplementary Fig. 17). Furthermore, CFD simulations were conducted to understand the underlying aerosol-based ink-mixing mechanisms under different sheath gas flow conditions. CFD analyses, combined with mixing index calculations, show considerable enhancement in the mixing of two aerosol ink flows when increasing sheath gas flow rates (for CFD model see Supplementary

Fig. 20; for ink mixing versus sheath flow, Supplementary Fig. 21; for mixing index and flow profile, Supplementary Figs. 22 and 23; and for simulation parameters, Supplementary Table 6), which is consistent with our experimental observations from fast camera imaging. This is an important finding because it provides a previously unexplored avenue to dynamically control ink mixing on the fly using sheath gas flow-induced aerodynamic focusing without the need for a complicated external mixer. In addition, CFD results indicate that a smaller nozzle diameter can enhance the convergence and mixing of aerosols (Supplementary Fig. 24), although an overly small nozzle (under 50 μm) may increase the probability of clogging during the printing process.

**Fig. 2: Rapid printing of combinatorial materials with gradient compositions.**




**a**, Fast camera image showing the aerosol deposition process of photoluminescent inks (blue) under visible light and ultraviolet light of 365 nm wavelength. Scale bar, 2.5 mm. **b**, Typical gradient profile of the HTCP

process. SEM images demonstrate the morphology changes in Ag/$Bi_2Te_3$ composites with increasing Ag ink flow and decreasing $Bi_2Te_3$ ink flow. The fabrication process takes approximately 1 min. Scale bars, 1 µm. **c**, Compositional characterization of Ag/$Bi_2Te_3$ composites with SEM (top) and EDS (bottom). Scale bars, 100 µm. **d**, Elemental distributions of Ag/$Bi_2Te_3$ composites over gradient direction. Error bars represent s.d. from three gradient samples. **e**, TEM of Ag/$Bi_2Te_3$ composites showing the interface between Ag nanoparticles and $Bi_2Te_3$ nanoplates. Scale bar, 2 nm.

As a demonstration of the HTCP process, we printed a metal/semiconductor nanocomposite using two inks containing zero-dimensional (0D) Ag nanoparticles of around 60 nm in diameter and two-dimensional (2D) $Bi_2Te_3$ nanoplates of around 1 µm in lateral size. Once a narrow stream of mixed aerosols was formed by the optimized aerodynamic focusing, we gradually increased the metal-to-semiconductor ink-mixing ratio and observed a clear morphological transition from a $Bi_2Te_3$ nanoplate-rich phase to a well-blended composite phase and then to a Ag nanosphere-rich phase (Fig. 2b). Changes in chemical composition were confirmed by energy-dispersive X-ray spectroscopy (EDS), as shown in Fig. 2c. A clearly increasing trend of Ag content in the printed combinatorial film was observed (Fig. 2d). Similar to the cosputtering method[1], HTCP is not aimed at generation of strictly linear-gradient compositions in combinatorial materials; instead, it is intended to produce gradient sample features in a fast, monotonic and high-throughput manner. Whereas certain process variations exist in HTCP printing, as indicated by the error bars in EDS measurements, a monotonic compositional gradient and distinct compositions of roughly 25 µm spatial resolution are observed along the gradient printing direction (elemental distribution, Fig. 2d; sample-to-sample reproducibility, Supplementary Fig. 18; transmission electron microscopy (TEM) of Ag/$Bi_2Te_3$ showing chemically fused nanocomposites after thermal sintering, Fig. 2e; TEM–EDS analysis, Supplementary Fig. 19).

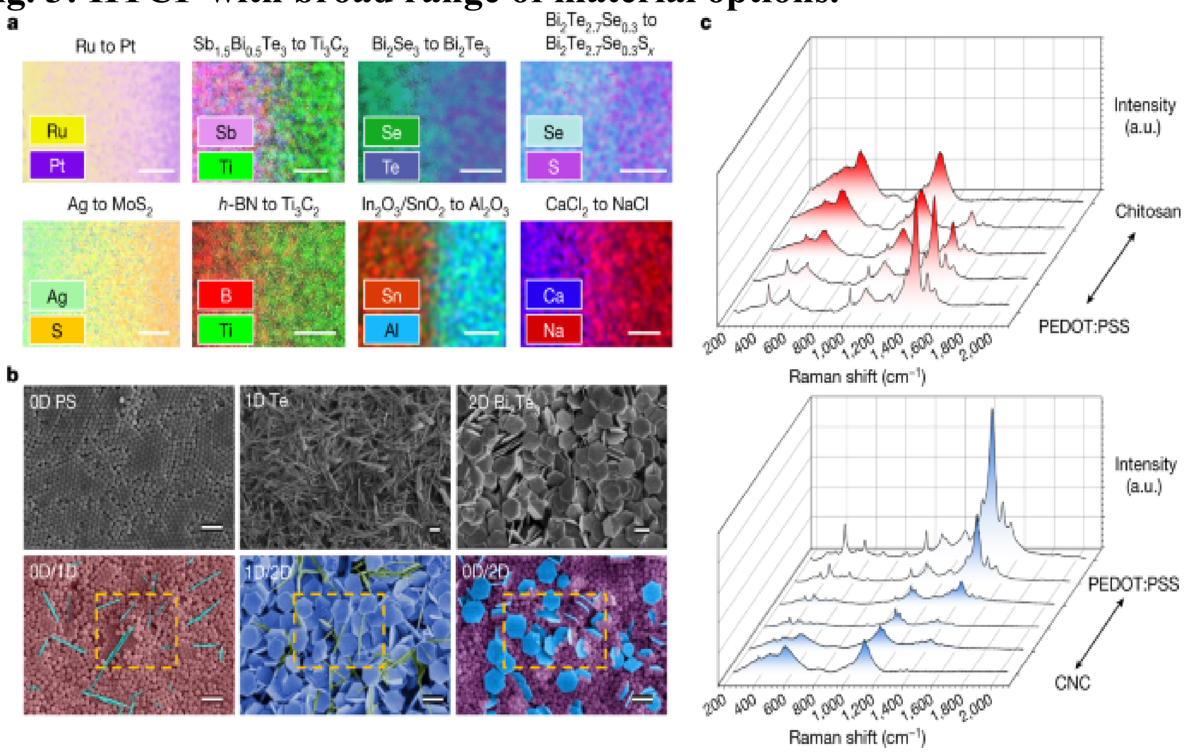
In general, composite fabrications involve a process that mixes one or several filler materials into a matrix material to achieve synergistic properties. A conventional trial-and-error approach often requires extensive processing time, which not only causes difficulty with high-throughput

fabrication but also may lead to undesired side reactions from the mismatch of starting materials related to their surface charge, pH values and ionic strength[21,22]. For example, MXene and $Sb_2Te_3$ nanoparticles exhibit opposite surface charges in a certain pH range[23], leading to the formation of large aggregates with poor colloidal stability (Supplementary Fig. 25). By contrast, the HTCP technique enables the rapid fabrication of combinatorial samples with gradient compositions, which minimizes undesired side effects (for example, aggregation). Aerosol-based HTCP uses nitrogen as the carrying gas, which forms a protective 'gas barrier' between ink droplets such that MXene and $Sb_2Te_3$ will not interact/react until being deposited onto the desired location of substrates. Consequently a dense, uniform composite film of MXene/$Sb_2Te_3$ was successfully printed (Supplementary Fig. 26). Thus, the combinatorial printing of seemingly incompatible materials is particularly unique and different from previous printing methods.

To explore the full capacity of the HTCP method, we fabricated a wide spectrum of films with gradient compositions including metals, oxides, nitrides, carbides, chalcogenides and halides, containing elements from s-block (groups IA–IIA) and p-block (groups IIIA–VIIA) of the periodic table (Fig. 3a). We also printed inks containing several d-block elements into combinatorial material libraries, including transition metal chalcogenides (for example, $MoS_2$) and transition metal carbides (for example, MXenes) (Supplementary Fig. 27). Moreover, the HTCP approach also exhibits excellent tolerance to material dimensions and morphology, as demonstrated in a printed 0D/1D composite of polystyrene (PS)/Te nanowires, a 1D/2D composite of Te nanowires/$Bi_2Te_3$ nanoplates and a 0D/2D composite of PS/$Bi_2Te_3$ nanoplates (scanning electron microscopy (SEM) images, Fig. 3b). SEM analyses on fractured cross sections of 0D/2D composite of Ag/$Bi_2Te_3$ reveal a homogeneous distribution of nanoscale building blocks without phase separations (Supplementary Fig. 28). In addition to inorganic nanomaterials, temperature-sensitive polymers (including biopolymers and semiconducting polymers) were tested (Fig. 3c). The Raman spectra of combinatorial polymer films of chitosan and poly(3,4-ethylenedioxythiophene)-poly(styrenesulfonate) (PEDOT:PSS) show a peak evolution at approximately 1,094 $cm^{-1}$ for chitosan[24] and at approximately

1,434 cm$^{-1}$ for PEDOT:PSS[25], indicating a clear compositional transition from chitosan to PEDOT:PSS. Similarly, cellulose nanocrystals (CNC) and PEDOT:PSS were printed into a gradient material library and the compositionally varying feature was verified using Raman analysis. These results demonstrate the versatility of HTCP in the rapid fabrication of a broad range of inorganic and organic combinatorial materials, substantially expanding the material options for high-throughput additive manufacturing.

**Fig. 3: HTCP with broad range of material options.**



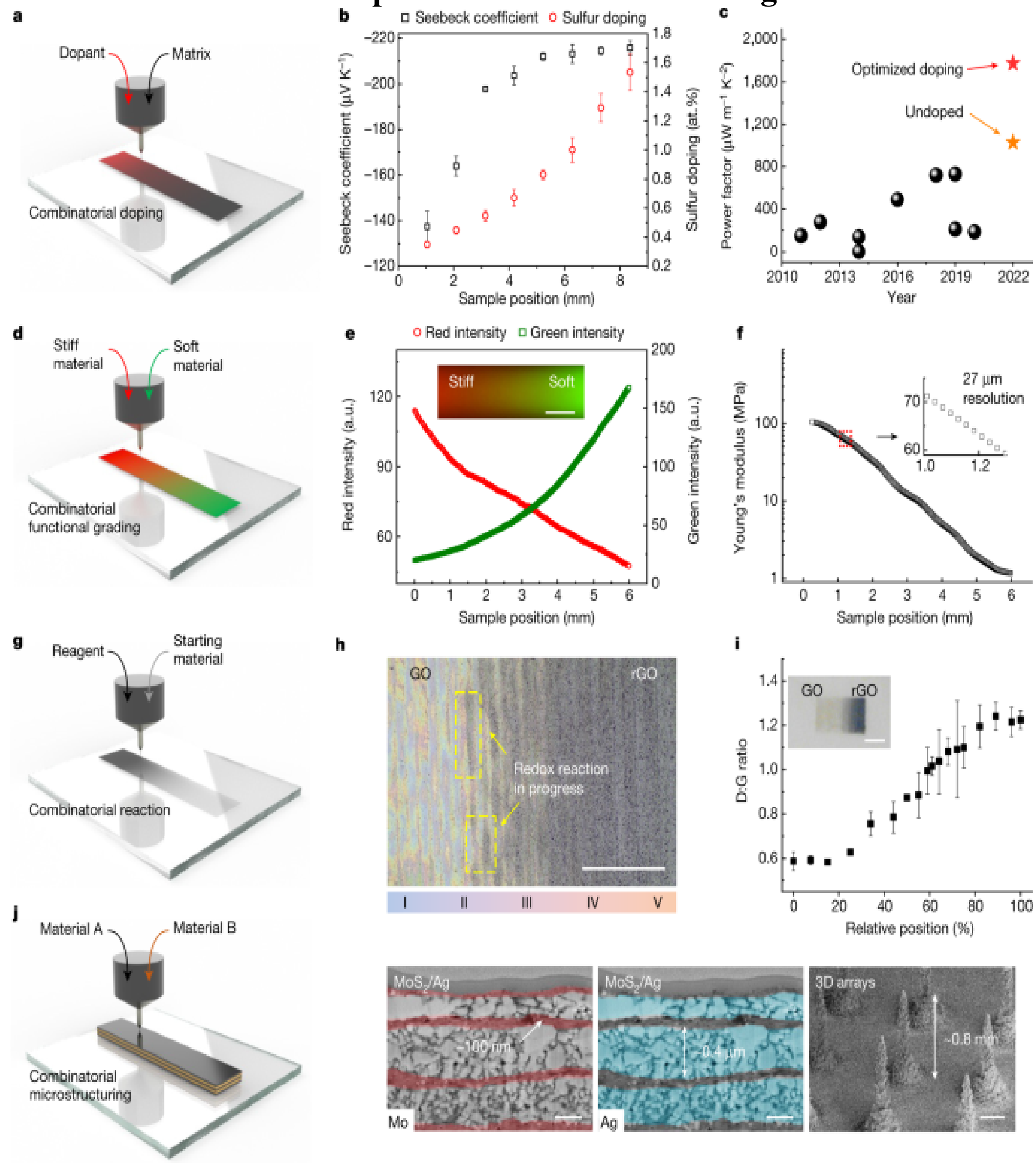

**a**, Elemental mapping of various printed combinatorial systems covering a broad range of elements. Scale bars, 300 µm. **b**, SEM images of various combinatorial materials printed directly using 0D, 1D and 2D nanoparticles. Scale bars, 1 µm. **c**, Combinatorial printing of polymers. Raman spectra confirm the compositional shift in combinatorial PEDOT:PSS/chitosan (top) and CNC/PEDOT:PSS (bottom). a.u., arbitrary units.

# HTCP of functional systems

To explore the potential of using the printed material library to accelerate materials screening and optimization toward the desired properties, we

demonstrated a high-throughput combinatorial doping strategy for thermoelectric applications (Fig. 4a). Due to high scalability and design freedom, thermoelectric (TE) printing has been extensively pursued in the past decade in the development of flexible or 3D conformal devices for energy harvesting and cooling. However, the relatively low performance (typically $10^2$–$10^3$ μW m$^{-1}$ K$^{-2}$ in power factor[26,27,28,29,30,31,32]) of printed n-type materials has thwarted the prospect of realizing broad impacts of printed thermoelectrics. To improve printed n-type materials, HTCP was used for rapid optimization of sulfur doping concentrations in printed $Bi_2Te_{2.7}Se_{0.3}$ materials, where samples of gradient doping concentrations were printed and tested in a single combinatorial gradient film. With increasing sulfur doping concentration, the Seebeck coefficient of printed $Bi_2Te_{2.7}Se_{0.3}$ film sharply increased from −130 to −200 μV K$^{-1}$ (at about 0.5% S) and then reached a plateau at around −213 μV K$^{-1}$ (at about 1.0% S; Fig. 4b). Such profound change could have originated from the increased density of states (DOS) effective mass \({m}_{{\rm{DOS}}}^{* }\) enabled by sulfur doping[33]. The combinatorial doping of a $Bi_2Te_{2.7}Se_{0.3}$ film shows a peak thin-film power factor at an optimal sulfur doping concentration of about 1.0 atomic weight percentage (at.%) (TE property versus doping; Supplementary Fig. 29). This printed material library is primarily aimed at identification of the optimized doping composition as opposed to achieving absolute property values. Because the electrical conductivity of the aerosol jet-printed films is limited by the requirement of using low-viscosity inks with relatively low particle concentration, an extrusion printing technique was applied to convert highly concentrated $Bi_2Te_{2.7}Se_{0.3}$ inks to thick films for practical device applications. Subsequent TE property measurements show a maximum room temperature power factor of 1,774 μW m$^{-1}$ K$^{-2}$ at 1.0% sulfur doping (Supplementary Fig. 30), which is substantially higher than most printed n-type TE materials (Fig. 4c and Supplementary Table 7)[26,27,28,29,30,31,32]. From a fundamental perspective, we also explored the combinatorial feature of HTCP to understand the compositional effect on Seebeck coefficient and charge carrier transport behaviour (ternary $Sb_xBi_{2-x}Te_3$ and quaternary $Sb_xBi_{(0.3x+6.7y)}Te_{(2x+9y)}Se_y$ alloys; Supplementary Fig. 31). These results demonstrate the data-rich feature of HTCP in efficient identification of optimized material compositions for achieving the desired properties.

**Fig. 4: HTCP enables combinatorial doping, functional grading, chemical reaction and compositional microstructuring.**




**a**, Schematic of combinatorial doping. **b**, A $Bi_2Te_{2.7}Se_{0.3}$ film with gradient sulfur doping concentrations and resulting local Seebeck coefficient changes. Error bars represent s.d. from two experimental replicates for Seebeck coefficient and six experimental replicates for sulfur doping concentration. **c**, Room temperature power factor comparisons between our

printed $Bi_2Te_{2.7}Se_{0.3}$ with optimized doping (red) and other printed n-type materials over the past decade[26,27,28,29,30,31,32], where our undoped $Bi_2Te_{2.7}Se_{0.3}$ is coloured orange for reference. **d**, Schematic of combinatorial printing of functionally graded materials. **e**, Fluorescent image showing red and green light intensities of a gradient polyurethane film printed with two PUD inks coloured by two dyes, with the inset showing a photoluminescent image. Scale bar, 1 mm. **f**, Local Young's modulus versus sample position measured with spatial resolution of about 27 µm. **g**, Schematic of combinatorial chemical reaction. **h**, Optical microscopic image of a GO/rGO gradient film. Scale bar, 0.5 mm. **i**, Raman analysis showing the D:G band ratio at different sample locations, with the inset showing an optical image of a gradient film following overnight reaction. Scale bar, 1 mm. Error bars represent s.d. from three experimental replicates. **j**, Left, schematic of combinatorial microstructuring. Right, examples of $Ag/MoS_2$ with periodic structures of submicron resolution along the thickness direction (false-coloured SEM showing Mo in red and Ag in blue; scale bars, 300 nm) and printed 3D arrays of salt crystals (scale bar, 300 µm).

In addition to material screening, we explored the potential of HTCP in the fabrication of functionally graded materials. As proof of concept, we printed gradient polyurethane films using two polyurethane dispersion (PUD) inks with different elastic moduli (Fig. 4d). The resultant gradient mixing can be visualized via fluorescent imaging by incorporation of the two PUD inks with red and green fluorescent dye, respectively (Fig. 4e, inset). A subsequent red, green and blue analysis of the functionally graded polyurethane (FGP) shows a monotonic trend of the compositional gradient with fluorescent intensity measured every 22.8 µm over the entire 6-mm-long film (Fig. 4e). To measure the mechanical property of the FGP, a tensile test accompanied by the 2D digital image correlation (DIC) method was used to map the strain field, and thereby to obtain the distribution of Young's modulus with a spatial resolution of around 27 µm. As shown in Fig. 4f, with increasing mixing ratio of soft to stiff PUD inks, FGP shows a monotonic decrease in Young's modulus over two orders of magnitude (from 103 to 1.2 MPa). Materials with such a gradient modulus may cover a range of biomaterials (tendon, skin, muscle and so on)[34] and may find applications in such as interface materials between components of different mechanical properties (Young's modulus versus biomaterials,

Supplementary Fig. 32; stretchability test, Supplementary Fig. 33). These results indicate the ability of HTCP to achieve monotonic gradation of both compositions and properties at 20–30 μm spatial resolution.

Besides printing stable inks, we investigated HTCP using reactive inks and their combinatorial reaction behaviours (Fig. 4g). During this process, reactive ink materials may undergo chemical/biochemical reactions triggered by the convergence of two inks and/or stimuli such as light, heat or catalysts. As proof of concept, graphene oxides (GOs) were coprinted with ascorbic acid (AC) under a gradient mixing ratio in which AC reduces GO to reduced graphene oxide (rGO). As the reaction proceeds, it is evident that a higher AC concentration causes a more rapid change in GO colour, from light yellow to dark brown, in the AC-rich region (Fig. 4h). Once the gradient-reduction reaction had reached completion, the inset of Fig. 4i shows the combinatorial film with a gradient appearance from light yellow (GO) to black (rGO). Raman spectroscopy (Fig. 4i and Supplementary Fig. 34) shows a clear change in D:G band ratio with increasing AC ink flow. This indicates that the reduction in GO decreased the average size of the $sp^2$ domains, because new graphitic domains were created with smaller sizes and larger quantities compared with those in the unreduced GO[35]. The HTCP method can also enable heterogenous fabrication of materials with compositionally complex structures by alternately depositing two ink materials layer by layer, leading to combinatorial microstructuring (Fig. 4j). An advantage of aerosol-based ink deposition is the ability to quickly switch from one material to the other due to the low viscous drag (for example, the $Ag/MoS_2$ nanocomposite with periodic structures shown in Fig. 4j), achieving submicron spatial resolution (roughly 100 nm) along film thickness direction, which is difficult to realize using other multimaterial printing methods (for example, extrusion printing). The Ag and $MoS_2$ inks for compositional microstructuring were formulated to be immiscible, to minimize diffusion between different layers. The microstructure patterning can also be applied to achieve 3D structures with a high aspect ratio (for example, the 3D pillars shown in Fig. 4j).

# Conclusions

The HTCP method enables high-throughput fabrication of versatile material libraries with gradient compositions utilizing rapid aerosol-based mixing and modulation of the mixing ratio. This in situ mixing and printing method may spark multiple potential research directions. First, HTCP can fabricate gradient films of metals, nitrides, carbides, chalcogenides, halides and even seemingly incompatible materials, enabling combinatorial materials screening and optimization with vastly expanded material options. Second, HTCP can produce functionally graded materials with unique compositional/structural arrangements and superior properties surpassing their constitutive materials with homogenous compositions. In addition, the combinatorial printing of reactive materials offers new possibilities toward high-throughput exploration, experimentation and characterization of chemicals/materials syntheses. The next phase of research will focus on leveraging the fabrication freedom and data-rich nature of HTCP, along with machine learning- and artificial intelligence-guided design strategies, which are expected to accelerate the discovery and development of a broad range of materials with intriguing and unprecedented properties for emerging applications.

# Methods

## Nanoparticle synthesis

Colloidal nanoparticles were prepared via either bottom-up synthesis or top-down exfoliation. For example, metal chalcogenides were synthesized via the bottom-up methods reported in our previous work[30,36]. Taking $Sb_2Te_3$ synthesis for example, a mixture of ethylene glycol and diethylene glycol in a proportion of 1:2, with a volume of 150 ml, was combined with 12 mmol of $SbCl_3$, 18 mmol of $TeO_2$, 3 g of NaOH and 0.8 g of polyvinylpyrrolidone with a molecular weight of 40,000 g $mol^{-1}$. The mixture was heated under reflux at 190 °C for 15 h and the precipitates separated by centrifugation at 5,000 rpm. To remove any remaining impurities, the precipitates were washed with ethanol three times. Similarly, $Bi_2Te_3$ synthesis was achieved using different metal precursors (for example, bismuth(III) nitrate pentahydrate). The $Ti_3C_2T_x$ MXene nanosheets (T represents surface termination, such as −O, −OH or −F) were synthesized by selective etching

of aluminium from the MXene precursor, $Ti_3AlC_2$, using a combined hydrofluoric acid/hydrochloric acid mixture and lithium chloride for delamination, as described previously[37]. For carbon-based nanosheets, liquid-phase exfoliation methods were used. Additional details on material synthesis are available in the Supplementary Information.

## General ink formulation

In a typical aqueous ink formulation a mixed solvent of water and ethylene glycol was used to disperse nanoparticles, in which ethylene glycol serves as a cosolvent to improve ink stability and printability. Depending on the ink type, a small amount of isopropyl alcohol may also be added as a defoamer to suppress the formation of foam during the ultrasonic atomizing process. To avoid aggregation and ensure uniform dispersion of the inks, they were sonicated for 15 min (Hilsonic bath sonicator, 300 W). Typical ink composition can be found in Supplementary Table 2. The concentration of nanomaterial was determined by the drying weight method. Commercially available inks, including indium tin oxides (30 wt.% in isopropanol), molybdenum disulfide (in terpineol/cyclohexanone) and aluminium oxide nanoparticle (20 wt.% in $H_2O$), were purchased from Sigma-Aldrich. Silver nanoparticle ink (PRELECT TPS 50G2) was obtained from Clariant Specialty Chemicals. For nonparticle inks, molecular or salt precursors were directly dissolved in solvents before the printing processes.

## HTCP printing

The motion control graphical user interface (GUI) controls the *x*, *y*, *z* motion stages of the printer. Real-time position and velocity were tracked for *x*, *y*, *z* during the targeted move and jogging operations. Both aerosol ink flow rates were actively controlled with respect to stage position to achieve gradient and voxelated films. For all high-resolution printing, the inks were first primed under sonication for 30 min and then atomized via ultrasonication before transfer to the printhead. A sheath flow was used to focus the aerosolized ink stream to achieve high printing resolution. Before printing, substrates (such as glass, mica, Kapton and so on) were precleaned with isopropyl alcohol and treated with plasma to improve surface hydrophilicity. During the HTCP process, a heating stage was used to evaporate the ink

solvents to minimize undesired drying effects. Depending on the type of combinatorial materials, additional thermal sintering was adopted to achieve the desired microstructures and properties.

## Thermoelectric property characterization

The Seebeck coefficient and electrical conductivity were measured using a custom-built scanning probing system consisting of two fine 40 AWG k-type thermocouples placed about 1 mm apart and a heater placed about 1 mm away from one of the thermocouples. Two electrodes through which current was sent for electrical conductivity measurement were placed at both ends of the printed film. At each measurement location, electrical conductivity was measured based on a linear four-probe principle at thermal equilibrium before measuring the Seebeck coefficient. During Seebeck coefficient measurement, heating power was slowly increased and Seebeck voltage between the two thermocouples was collected continuously along with thermocouple temperatures. The absolute Seebeck coefficient of the film was calculated by taking into account the Seebeck coefficient of the thermocouple wire. The measurement system was validated by measuring a bulk constantan film with known thermoelectric properties. In addition, the printed gradient films were also measured using a custom-built scanning probe instrument at the National Institute of Standards and Technology[38]. The Seebeck coefficient was measured in 0.5 mm increments using a pair of independently spring-loaded type R thermocouple probes spaced 3 mm apart, using the quasisteady state condition of the differential method[39]. Measurement uncertainty (1 s.d.) for Seebeck measurement is ±6.5%. Details of these thermoelectric measurements can be found in Supplementary Information.

## Mechanical property mapping of gradient polyurethane films

To obtain Young's modulus distribution of a printed gradient polyurethane film, we used the 2D DIC method to track strain and displacement distributions under uniaxial stretch. The specimen had a length of 12 mm, width of 2 mm and a gradient composition. The force–displacement relation was recorded by uniaxial stretching of the specimen up to 50% strain at a rate of 0.1% $s^{-1}$ via an Instron universal machine (model 5944) with a 50 N

load cell. The specimen was mounted in a pair of tensile grips, leaving a gauge length of 8 mm. The ratio of length to width (4) is sufficiently high to ensure that the majority of the specimen undergoes uniaxial tension with negligible edge effects. To measure strain distribution by the DIC method we sprayed ink (Koh-I-Noor Rapidraw) with an airbrush (Badger, no. 150) to generate high-quality speckle patterns on the specimen. To enhance optical contrast, a whiteboard was used as background and a white light LED light was shot on the sample during testing. Changes in speckle patterns were recorded by a Canon ESO 6D DSLR camera with a Canon 100 mm F/2.8L macro lens at roughly every 1% strain. The resolution of each image was around 6.8 μm per pixel, and we output data every four-pixel length. Images were analysed by Ncorr[40], an open-source 2D DIC Matlab software, to obtain the strain and displacement distributions of the middle region of 6 mm (L) × 2 mm (W). We tested the specimen four times to eliminate natural error and, after each loading and unloading, the specimen was placed on a hotplate at 35 °C for 5 min, and at room temperature for a further 10 min, to fully release any residual stress. Based on the measured strain distribution, we calculated Young's modulus distribution in the elongation direction under the assumption of linearly elastic material and no gradient modulus along the width (Supplementary Information).

## Data availability

The datasets generated or analysed during the current study are available from the corresponding author on request.

## References

1. Ludwig, A. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *NPJ Comput. Mater.* **5**, 70 (2019).

2. McGinn, P. J. Thin-film processing routes for combinatorial materials investigations—a review. *ACS Comb. Sci.* **21**, 501–515 (2019).

3. Xiang, X.-D. et al. A combinatorial approach to materials discovery. *Science* **268**, 1738–1740 (1995).

4. Lopes, L. R., Silva, A. F. & Carneiro, O. S. Multi-material 3D printing: the relevance of materials affinity on the boundary interface performance. *Addit. Manuf.* **23**, 45–52 (2018).

5. Skylar-Scott, M. A., Mueller, J., Visser, C. W. & Lewis, J. A. Voxelated soft matter via multimaterial multinozzle 3D printing. *Nature* **575**, 330–335 (2019).

6. Xing, X. et al. High-resolution combinatorial patterning of functional nanoparticles. *Nat. Commun.* **11**, 6002 (2020).

7. Green, M. L., Takeuchi, I. & Hattrick-Simpers, J. R. Applications of high throughput (combinatorial) methodologies to electronic, magnetic, optical, and energy-related materials. *J. Appl. Phys.* **113**, 231101 (2013).

8. Jiang, C. et al. Printed subthreshold organic transistors operating at high gain and ultralow power. *Science* **363**, 719–723 (2019).

9. Kim, Y., Yuk, H., Zhao, R., Chester, S. A. & Zhao, X. Printing ferromagnetic domains for untethered fast-transforming soft materials. *Nature* **558**, 274–279 (2018).

10. Saccone, M. A., Gallivan, R. A., Narita, K., Yee, D. W. & Greer, J. R. Additive manufacturing of micro-architected metals via hydrogel infusion. *Nature* **612**, 685–690 (2022).

11. Chen, X. et al. Multi-metal 4D printing with a desktop electrochemical 3D printer. *Sci. Rep.* **9**, 3973 (2019).

12. Reiser, A. et al. Multi-metal electrohydrodynamic redox 3D printing at the submicron scale. *Nat. Commun.* **10**, 1853 (2019).

13. Queraltó, A. et al. Combinatorial screening of cuprate superconductors by drop-on-demand inkjet printing. *ACS Appl. Mater. Interfaces* **13**, 9101–9112 (2021).

14. Wang, K., Chang, Y.-H., Zhang, C. & Wang, B. Conductive-on-demand: tailorable polyimide/carbon nanotube nanocomposite thin film by dual-material aerosol jet printing. *Carbon* **98**, 397–403 (2016).

15. Craton, M. T., Albrecht, J. D., Chahal, P. & Papapolymerou, J. Multimaterial aerosol jet printed magnetic nanocomposites for microwave circuits. *IEEE Trans. Compon. Packaging Manuf. Technol* **11**, 865–871 (2021).

16. Mahajan, A., Frisbie, C. D. & Francis, L. F. Optimization of aerosol jet printing for high-resolution, high-aspect ratio silver lines. *ACS Appl. Mater. Interfaces* **5**, 4856–4864 (2013).

17. Tafoya, R. R. et al. Real-time optical process monitoring for structure and property control of aerosol jet printed functional materials. *Adv. Mater. Technol.* **5**, 2000781 (2020).

18. Zhang, H., Moon, S. K. & Ngo, T. H. Hybrid machine learning method to determine the optimal operating process window in aerosol jet 3D printing. *ACS Appl. Mater. Interfaces* **11**, 17994–18003 (2019).

19. Zeng, M. et al. Scalable nanomanufacturing of chalcogenide inks: a case study on thermoelectric V–VI nanoplates. *J. Mater. Chem. A* **9**, 22555–22562 (2021).

20. Zeng, M. et al. Colloidal nanosurfactants for 3D conformal printing of 2D van der Waals materials. *Adv. Mater.* **32**, e2003081 (2020).

21. Bouyer, F., Robben, A., Yu, W. L. & Borkovec, M. Aggregation of colloidal particles in the presence of oppositely charged polyelectrolytes: effect of surface charge heterogeneities. *Langmuir* **17**, 5225–5231 (2001).

22. Lin, W. et al. Heteroaggregation in binary mixtures of oppositely charged colloidal particles. *Langmuir* **22**, 1038–1047 (2006).

23. Lu, X. et al. High-efficiency thermoelectric power generation enabled by homogeneous incorporation of MXene in (Bi,Sb)2Te3 matrix. *Adv. Energy Mater.* **10**, 1902986 (2020).

24. Souza, N. L. G. D., Brandão, H. M. & de Oliveira, L. F. C. Spectroscopic and thermogravimetric study of chitosan after incubation in bovine rumen. *J. Mol. Struct.* **1005**, 186–191 (2011).

25. Lee, H. J. et al. Enhanced thermoelectric performance of PEDOT:PSS/PANI–CSA polymer multilayer structures. *Energy Environ. Sci.* **9**, 2806–2811 (2016).

26. We, J. H., Kim, S. J. & Cho, B. J. Hybrid composite of screen-printed inorganic thermoelectric film and organic conducting polymer for flexible thermoelectric power generator. *Energy* **73**, 506–512 (2014).

27. Chen, A., Madan, D., Wright, P. K. & Evans, J. W. Dispenser-printed planar thick-film thermoelectric energy generators. *J. Micromech. Microeng.* **21**, 104006 (2011).

28. Madan, D. et al. Enhanced performance of dispenser printed MA n-type Bi2Te3 composite thermoelectric generators. *ACS Appl. Mater. Interfaces* **4**, 6117–6124 (2012).

29. Kim, F. et al. 3D printing of shape-conformable thermoelectric materials using all-inorganic Bi2Te3-based inks. *Nat. Energy* **3**, 301–309 (2018).

30. Saeidi-Javash, M., Kuang, W., Dun, C. & Zhang, Y. 3D conformal printing and photonic sintering of high-performance flexible thermoelectric films using 2D nanoplates. *Adv. Funct. Mater.* **29**, 1901930 (2019).

31. Ferhat, S. et al. Flexible thermoelectric device based on TiS2(HA)x n-type nanocomposite printed on paper. *Org. Electron.* **68**, 256–263 (2019).

32. Su, N., Zhu, P., Pan, Y., Li, F. & Li, B. 3D-printing of shape-controllable thermoelectric devices with enhanced output performance. *Energy* **195**, 116892 (2020).

33. Devender, et al. Harnessing topological band effects in bismuth telluride selenide for large enhancements in thermoelectric properties

through isovalent doping. *Adv. Mater.* **28**, 6436–6441 (2016).

34. McKee, C. T., Last, J. A., Russell, P. & Murphy, C. J. Indentation versus tensile measurements of Young's modulus for soft biological tissues. *Tissue Eng. Part B Rev.* **17**, 155–164 (2011).

35. Stankovich, S. et al. Synthesis of graphene-based nanosheets via chemical reduction of exfoliated graphite oxide. *Carbon* **45**, 1558–1565 (2007).

36. Dun, C. et al. 3D printing of solution-processable 2D nanoplates and 1D nanorods for flexible thermoelectrics with ultrahigh power factor at low-medium temperatures. *Adv. Sci.* **6**, 1901788 (2019).

37. Saeidi-Javash, M. et al. All-printed MXene–graphene nanosheet-based bimodal sensors for simultaneous strain and temperature sensing. *ACS Appl. Electron. Mater.* **3**, 2341–2348 (2021).

38. Yan, Y. G., Martin, J., Wong-Ng, W., Green, M. & Tang, X. F. A temperature dependent screening tool for high throughput thermoelectric characterization of combinatorial films. *Rev. Sci. Instrum.* **84**, 115110 (2013).

39. Martin, J. Protocols for the high temperature measurement of the Seebeck coefficient in thermoelectric materials. *Meas. Sci. Technol.* **24**, 085601 (2013).

40. Blaber, J., Adair, B. & Antoniou, A. Ncorr: open-source 2D digital image correlation Matlab software. *Exp. Mech.* **55**, 1105–1122 (2015).

# Acknowledgements

Y.Z. acknowledges funding support from the National Science Foundation (award no. CMMI-1747685), the US Department of Energy (award no. DE-EE0009103) and the Department of Energy Advanced Sensors and Instrumentation programme (DOE Idaho Operations Office, contract no. DE-AC07-05ID14517). D.J.K. acknowledges the support of the National Science Foundation Graduate Research Fellowship Program (grant no. DGE

1840340). C.W. and L.J. acknowledge funding support from the National Science Foundation (award no. CMMI-1925790). The authors acknowledge the use and support of Notre Dame Integrated Imaging Facility, Materials Characterization Facility and the Center for Environmental Science and Technology.

# Author information

Author notes

1. These authors contributed equally: Minxiang Zeng, Yipu Du, Qiang Jiang

## Authors and Affiliations

1. Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN, USA

   Minxiang Zeng, Yipu Du, Qiang Jiang, Nicholas Kempf, Miles V. Bimrose, A. N. M. Tanvir, Hengrui Xu, Jiahao Chen, Tatsunori Hayashi, Mortaza Saeidi-Javash, Hirotaka Sakaue & Yanliang Zhang

2. Department of Chemical Engineering, Texas Tech University, Lubbock, TX, USA

   Minxiang Zeng

3. Department of Mechanical and Aerospace Engineering, University of California Los Angeles, Los Angeles, CA, USA

   Chen Wei & Lihua Jin

4. Department of Mechanical Science and Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA

   Miles V. Bimrose

5. Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

   Dylan J. Kirsch & Joshua Martin

6. Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA

   Dylan J. Kirsch

7. Department of Mechanical and Energy Engineering and Integrated Nanosystems Development Institute, Purdue School of Engineering and Technology, Indiana University−Purdue University Indianapolis, Indianapolis, IN, USA

   Brian C. Wyatt & Babak Anasori

8. Department of Mechanical and Aerospace Engineering, California State University Long Beach, Long Beach, CA, USA

   Mortaza Saeidi-Javash

9. Idaho National Laboratory, Idaho Falls, ID, USA

   Michael D. McMurtrey

## Contributions

Y.Z. developed the HTCP concept. M.Z. and Y.Z. designed the overall experiments. Y.D. and M.V.B. developed software programmes for the HTCP process. Q.J. and A.N.M.T. contributed to the preparation and characterization of combinatorial samples and thermoelectric samples as well as statistical analysis. N.K., D.J.K. and J.M. carried out transport property measurements. B.C.W. and B.A. prepared MXene nanoparticles. T.H. and H.S. performed fast camera experiments. H.X. contributed to CFD simulation. M.S.-J. contributed to sintering experiments. J.C. and M.D.M. contributed to noble metal nanoparticle printing experiments. Y.D., Q.J., C.W. and L.J. conducted polyurethane film printing and characterizations.

M.Z. and Y.Z. prepared the manuscript. All authors discussed the results and commented on the manuscript.

### Corresponding author

Correspondence to Yanliang Zhang.

## Ethics declarations

### Competing interests

The authors declare no competing interests. A provisional patent application (application no. 63/482,413) has been filed by University of Notre Dame on this research.

## Peer review

### Peer review information

*Nature* thanks Emilio Sardini and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

## Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Supplementary information

### Supplementary Information

Supplementary Information containing Methods, Figs. 1–34 and Tables 1–7; see Contents page for details.

### Supplementary Video 1

Demonstration of real-time HTCP process for the printing of gradient combinatorial materials, in which a gradient-coloured film was fabricated by shifting the mixing ratio of blue and yellow dye inks on the fly.

**Supplementary Video 2**

Typical aerosol jet-printing process for HTCP at different sheath flow rates. Despite the challenge of optical imaging on aerosol streams, it was observed that a higher sheath flow leads to a narrower aerosol stream. The video was taken using a high-speed camera under a frame rate of 5,000 $s^{-1}$.

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

## About this article

### Cite this article

Zeng, M., Du, Y., Jiang, Q. *et al.* High-throughput printing of combinatorial materials from aerosols. *Nature* **617**, 292–298 (2023). https://doi.org/10.1038/s41586-023-05898-9

- Received: 07 November 2021
- Accepted: 28 February 2023
- Published: 10 May 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05898-9

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

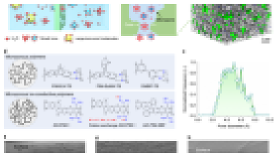
Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

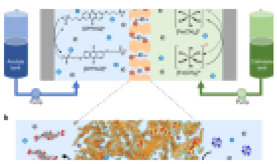
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05898-9



## Hydrophilic microporous membranes for selective ion separation and flow-battery energy storage

02 December 2019

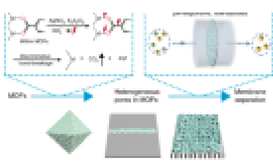
Rui Tan, Anqi Wang, … Qilei Song



## Development of efficient aqueous organic redox flow batteries using ion-sieving sulfonated polymer membranes

08 June 2022

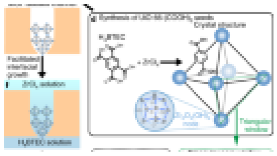
Chunchun Ye, Anqi Wang, … Qilei Song



## Metal-organic framework patterns and membranes with heterogeneous pores for flow-assisted switchable separations

27 September 2018

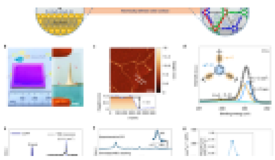
Guan-Young Jeong, Ajay K. Singh, … Dong-Pyo Kim



## Efficient metal ion sieving in rectifying subnanochannels enabled by metal–organic frameworks

09 March 2020

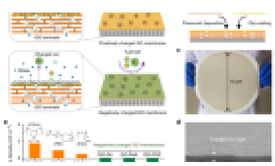
Jun Lu, Huacheng Zhang, … Huanting Wang



## Fast water transport and molecular sieving through ultrathin ordered conjugated-polymer-framework membranes

08 August 2022

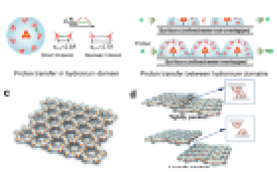
Jie Shen, Yichen Cai, … Yu Han



## Controllable ion transport by surface-charged graphene oxide membrane

19 March 2019

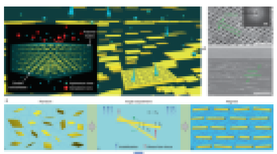
Mengchen Zhang, Kecheng Guan, … Nanping Xu



[Short hydrogen-bond network confined on COF surfaces enables ultrahigh proton conductivity]

05 November 2022

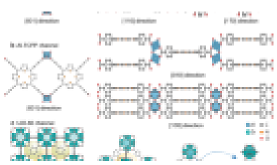
Benbing Shi, Xiao Pang, … Zhongyi Jiang



[Polymeric membranes with aligned zeolite nanosheets for sustainable energy storage]

17 October 2022

Yongsheng Xia, Hongyan Cao, … Zhi Xu



[Construction of angstrom-scale ion channels with versatile pore configurations and sizes by metal-organic frameworks]

18 January 2023

Xingya Li, Gengping Jiang, … Huacheng Zhang

- Article
- Published: 26 April 2023

# Near-frictionless ion transport within triazine framework

# membranes


- Peipei Zuo,
- Chunchun Ye,
- Zhongren Jiao,
- Jian Luo,
- Junkai Fang,
- Ulrich S. Schubert,
- Neil B. McKeown,
- T. Leo Liu ,
- Zhengjin Yang &
- …
- Tongwen Xu

Show authors


*Nature* **volume 617**, pages 299–305 (2023)


- 18k Accesses
- 80 Altmetric
- Metrics details


## Abstract

The enhancement of separation processes and electrochemical technologies such as water electrolysers[1,2], fuel cells[3,4], redox flow batteries[5,6] and ion-capture electrodialysis[7] depends on the development of low-resistance and high-selectivity ion-transport membranes. The transport of ions through these membranes depends on the overall energy barriers imposed by the collective interplay of pore architecture and pore–analyte interaction[8,9]. However, it remains challenging to design efficient, scaleable and low-cost selective ion-transport membranes that provide ion channels for low-energy-barrier transport. Here we pursue a strategy that allows the diffusion limit of ions in water to be approached for large-area, free-standing, synthetic

membranes using covalently bonded polymer frameworks with rigidity-confined ion channels. The near-frictionless ion flow is synergistically fulfilled by robust micropore confinement and multi-interaction between ion and membrane, which afford, for instance, a $Na^+$ diffusion coefficient of $1.18 \times 10^{-9}$ m$^2$ s$^{-1}$, close to the value in pure water at infinite dilution, and an area-specific membrane resistance as low as 0.17 Ω cm$^2$. We demonstrate highly efficient membranes in rapidly charging aqueous organic redox flow batteries that deliver both high energy efficiency and high-capacity utilization at extremely high current densities (up to 500 mA cm$^{-2}$), and also that avoid crossover-induced capacity decay. This membrane design concept may be broadly applicable to membranes for a wide range of electrochemical devices and for precise molecular separation.

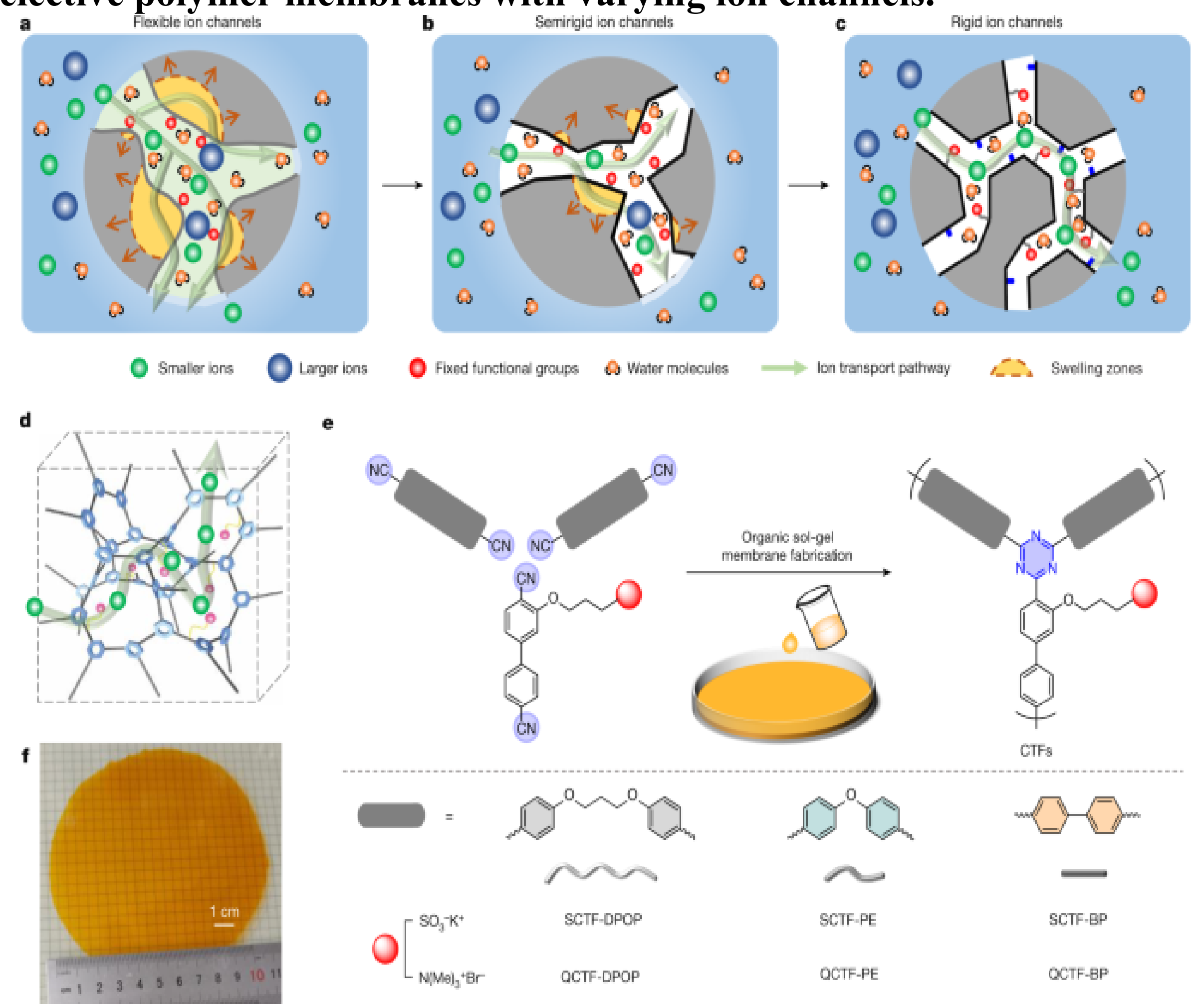
# Main

Selective ion-transport membranes are key components of clean-energy technologies, including large-scale, energy-efficient separation and purification processes and, of course, electrochemical devices as diverse as $CO_2$ electrolysers and water electrolysers[1,2,10], $H_2/O_2$ fuel cells[3,4], redox flow batteries[5,6], ion-capture electrodialysis[7] and so on. In all of these well-established and emerging electrochemical systems, membrane separators transport ions and isolate electrochemical reactions in two half-cells, their efficacy being contingent on the ability of membrane separators to perform rapid and selective ion transport[5,11,12]. Ion transport through membranes at subnanometre scales depends on the overall energy barriers, which are determined by nanoconfinement within free-volume voids and pore–ion interactions[8,9]. Therefore, construction of ion channels with low energy barriers is imperative for the development of high-performing membranes in electrochemical processes.

Polymer materials dominate the use of selective ion-transport membranes in practical modules across all scales, due to their ideal combination of low cost, manufacturing scaleability and small footprint. Conventional polymer materials, represented by perfluorocarbon Nafion and the recently developed hydrocarbon-based polyelectrolytes, form ion-conductive regions via microphase separation[13,14]. However, these regions are ill defined and prone

to swelling to several nanometres in size following hydration, which may facilitate ion diffusion but also causes poor selectivity (Fig. 1a and Supplementary Fig. 1). The emerging perspective is to confine the size of pore volume voids—that is, to increase the rigidity of the polymer backbone and reduce the thermal motion of the polymer segment—to enforce good size selectivity for rapid ion transport[15]. In particular, solution-processable, semirigid polymers of intrinsic microporosity (PIMs; Fig. 1b) are emerging as next-generation molecule-separation[16,17] and ion-transport membranes[18,19,20,21,22,23], benefiting from size-exclusion-induced selectivity and free-volume-induced permeability. However, these semirigid, non-network PIMs may age and, at high content of charged functionalities, can undergo severe swelling leading to decreased selectivity[21,24,25].

**Fig. 1: Schematic illustrations showing existing and proposed ion-selective polymer membranes with varying ion channels.**



**a**, Membranes with flexible ion channels. These contain microphase-separated morphology derived from the assembly of hydrophilic ion-conductive moieties and hydrophobic flexible-polymer backbones, represented by Nafion. **b**, Ion-selective microporous membranes with semirigid ion channels. The channels are formed by intrinsic micropores resulting from inefficient polymer chain packing, represented by polymers of intrinsic microporosity and their derivatives. To render the membrane ion conductive, functional moieties are incorporated during postsynthetic modification. Membranes may age over time and swell in water. **c**,**d**, Our proposed membranes with rigid ion channels (**c**). These are expected to build from bottom-up synthesis and via swelling-resistant 3D polymer frameworks (**d**). Pore architecture and chemistry are tuned for rapid and selective ion transport. **e**,**f**, Preparation of stand-alone CTF membranes via a superacid-catalysed organic sol-gel reaction from functional aromatic nitrile monomers (**e**). CTF membranes have a controlled number of ion-conductive moieties inside membrane pores and a covalent network structure. Image (**f**) shows a free-standing CTF membrane with a diameter of over 10 cm. Structure rigidity and microporosity of the CTF membrane can be regulated by designing variable structural units, as demonstrated at bottom right, from flexible to very rigid.

We therefore aimed to deploy shape-persistent polymer frameworks to apply rigid confinement to free-volume voids, but realized that a well-confined framework alone would be unlikely to provide a mechanism for rapid ion transport. Another key design criterion to reduce energy barriers of ion transport across membranes is to leverage interactions between pore walls and ions to promote frictionless ion transport. Thus, we sought to improve multiple interactions between pores and ions by engineering charged, polar and Lewis acidic/basic moieties into polymer frameworks. Here we report a new family of all-rigid polymer framework membranes containing well-confined ion channels (Fig. 1c,d). Our structural design combines (1) the isolation of absorbed water within microcavities from the hydrophobic backbone framework to facilitate selective ion transport at robust subnanometre confinement and (2) multiple interactions between ions and pore walls to synergistically approach the diffusion limit of ions based on coordination mechanisms. We exemplify this concept via the fabrication of covalent triazine framework (CTF) membranes whose channel size, as well as channel chemistry, can be finely-tuned. We find that these polymer

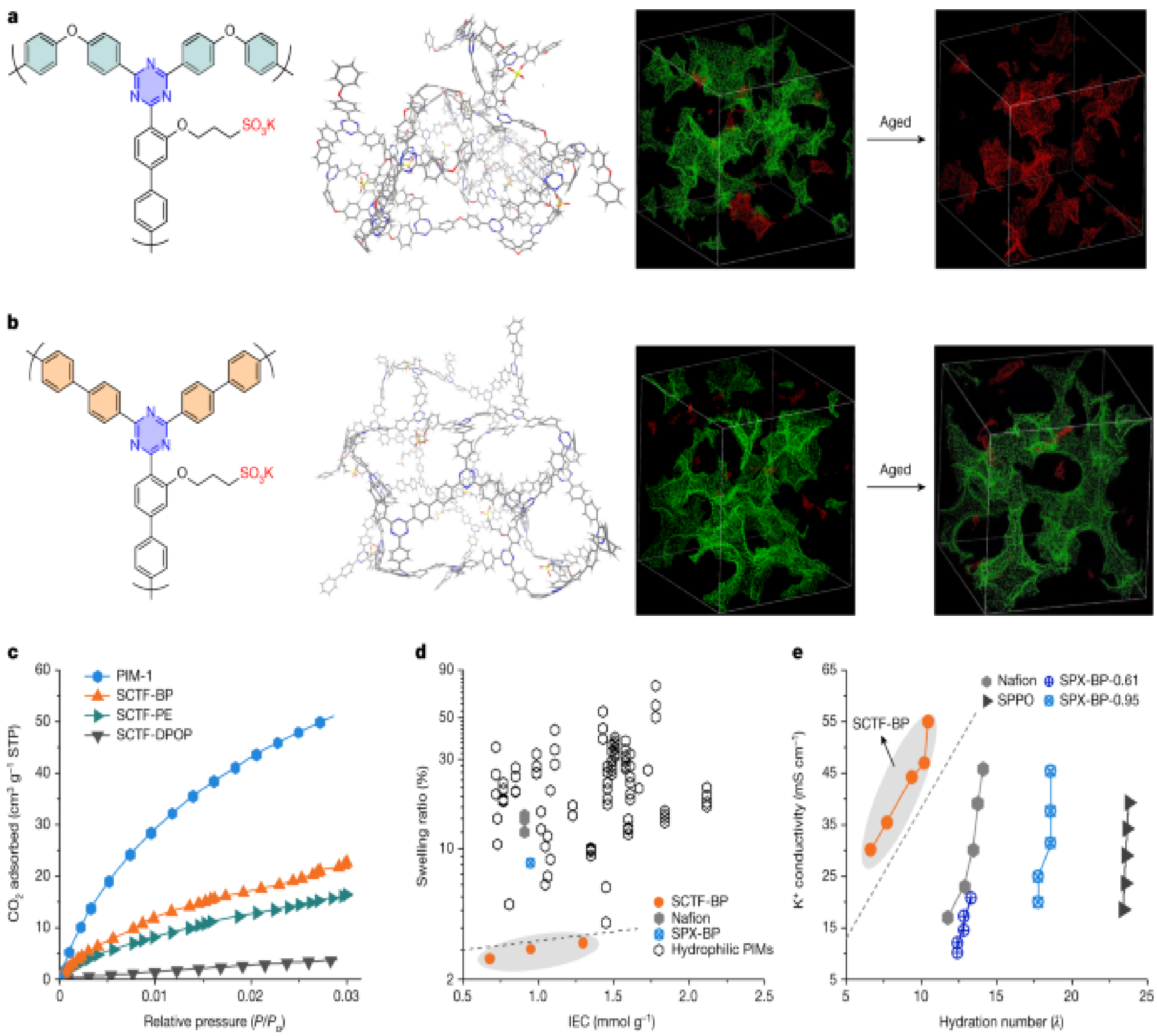
framework membranes exhibit ultrahigh ion diffusivity but also extremely low permeability of active materials. Using flow battery technology for exemplification, CTF membranes fulfilled their functions as efficient ion-transport membrane separators providing flow batteries with ultrafast charging capacity.

The CTF membranes are constructed from functional aromatic nitrile monomers via a superacid-catalysed organic sol-gel reaction[26,27] (Fig. 1e,f and Supplementary Figs. 2–4). Self-standing CTF membranes are readily formed on a large scale during the polymerization reaction with concomitant microporosity at the narrow subnanometre scale (Supplementary Fig. 5). Pore geometry is readily regulated by the appropriate choice of monomer structural units with varying rigidity; pore chemistry is adjusted by adopting different types of functional groups (that is, negatively, positively and neutrally charged moieties) and by controlling the amount of ion-conductive moieties (that is, ion-exchange capacity, IEC) to afford corresponding cation-conductive membranes (SCTFs) and anion-conductive membranes (Fig. 1e and Supplementary Fig. 3). Importantly short, flexible spacers are applied to connect nitrile monomers to charged functional groups, thereby mitigating their electron-withdrawing effects and enabling the trimerization reaction of aromatic nitriles. Details of CTF synthesis and spectroscopic characterization are provided in the Supplementary Information. The mechanical properties of the resulting polymer membranes were measured by tensile testing (Supplementary Fig. 5), showing Young's modulus and the ultimate tensile strength of SCTF-biphenyl (SCTF-BP) membranes as high as 1.1 GPa and 29 MPa, respectively.

Moderate $CO_2$ uptake of the SCTF-BP sample at low pressure and 273 K confirmed the accessible microporosity of the three-dimensional (3D) framework structure (Fig. 1d and Supplementary Fig. 6). Notably, SCTF-BP has lower $CO_2$ adsorption than classical semirigid PIM-1 (the archetypal PIM polymer), suggesting the smaller size of free-volume elements in SCTF-BP, and predominately subnanometre-sized pore size distribution was confirmed using positron annihilation lifetime spectroscopy (Supplementary Fig. 6), which would improve the size selectivity of ion transport. Micropore size distribution of the resulting membranes is regulated by adopting monomers of varying rigidity (Fig. 2a–c and Supplementary Fig. 7). SCTF-

BP constructed from rigid monomers can provide almost one order of magnitude higher cumulative pore volume in comparison with that of SCTF-diphenoxypropane (SCTF-DPOP) fabricated using flexible monomers, as visualized from molecular simulations of the accessible free-volume elements of SCTF-BP compared with other CTF membranes (Fig. 2a,b and Supplementary Figs. 8 and 9). After ageing these cell models with a forcefield, highly rigid SCTF-BP retained the most interconnected volume elements, suggesting its enhanced anti-swelling capacity (Fig. 2a,b and Supplementary Fig. 9). X-ray diffraction analysis showed that SCTF-BP has similar scattering peaks in dry and swelling states, indicating that interchain distance does not change following water uptake (Supplementary Fig. 6c, right). A control experiment performed on SPX-BP, a PIM with the same biphenyl linkage as SCTF-BP, showed the disappearance of distinct scattering peaks owing to the obvious swelling of the membrane (Supplementary Figs. 6c (left) and 10). Indeed, the SCTF-BP membrane has an extremely low swelling ratio in comparison with both commercial ion-exchange membranes and reported PIM membranes, in which high IEC values are correlated with severe swelling, thus compromising membrane selectivity (Fig. 2d and Supplementary Table 1).

**Fig. 2: Characterization of negatively charged CTF membranes (SCTF).**



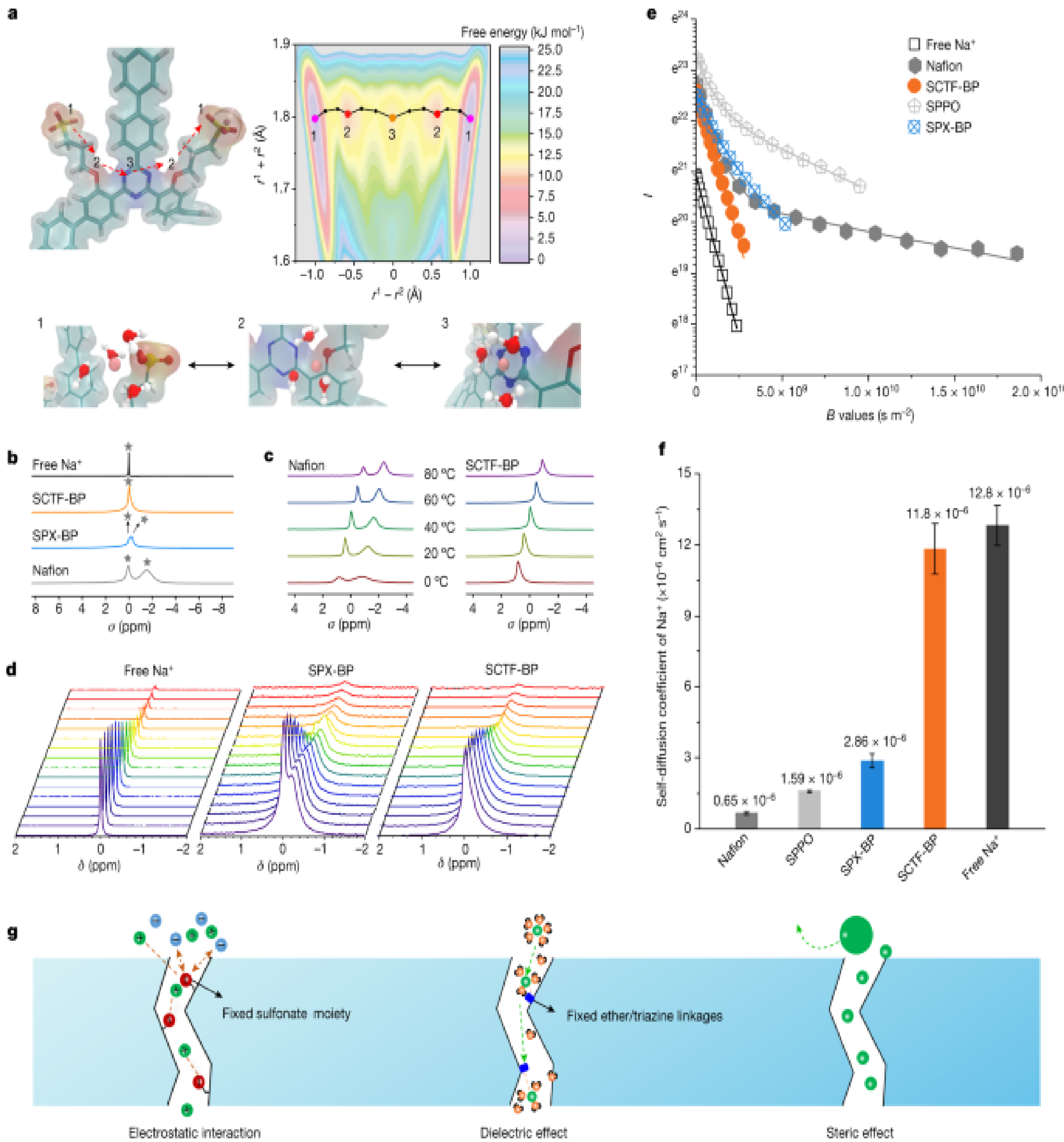

**a**,**b**, Schematic representation of the chain structure of SCTF-PE (**a**) and SCTF-BP (**b**), with a flexible and a very rigid chain, respectively (left). 3D view of the amorphous cells of SCTF-PE and SCTF-BP membranes (middle). Images of aged cells shown on the right. Red shading represents isolated free-volume elements and green shading represents interconnected micropores. **c**, $CO_2$ uptake (at 273 K) of SCTF samples compared with that of PIM-1. **d**, Swelling ratio plotted as a function of IEC (the content of charged functional groups, in mmol $g^{-1}$) for SCTF-BP, Nafion, SPX-BP and hydrophilic PIM membranes. **e**, $K^+$ conductivity plotted as a function of hydration number for SCTF-BP, Nafion, SPX-BP and SPPO membranes. Each dot represents measurement at a separate temperature: from left to right, 30–70 °C. Dashed lines and shading are a visual guide only. **d**,**e**, The

values shown can be found in Supplementary Tables 1 and 3. STP, standard temperature and pressure.

The robust framework structure with subnanometre micropores can effectively block the permeation of redox-active electrolytes through the SCTF-BP membrane. Over the temperature range of 25–60 °C, the permeability of 2,6-dihydroxyanthraquinone (DHAQ) is at least one order of magnitude lower, and that of $K_4[Fe(CN)_6]$ is two orders of magnitude lower through SCTF-BP, than Nafion 212 (Supplementary Figs. 11 and 12 and Supplementary Table 2; see Supplementary Information for pretreatment of Nafion membrane samples; note that Nafion properties may vary significantly with pretreatment). Such low permeabilities of active species, even at the high temperature of 60 °C, are envisioned as affording the SCTF-BP flow cell a decadal operational lifetime (Supplementary Fig. 12). Importantly, SCTF-BP membranes demonstrate rapid transport of charge carriers, with the most conductive $K^+$ showing a cation transference number ($t_+$) as high as 0.988 (Supplementary Fig. 13). The $K^+$ conductivity of the SCTF-BP membrane, as measured by four-point electrochemical impedance spectroscopy (EIS), approached 30 mS cm$^{-1}$ at 30.0 °C and 54.9 mS cm$^{-1}$ at 70 °C at low hydration number (6.7 at 30 °C, 10.5 at 70 °C; Fig. 2e, Supplementary Fig. 14 and Supplementary Table 3). Notably, for existing membranes with extremely low electrolyte permeability (for example, Fumasep E-620(K)), $K^+$ conductivity is far lower than that of SCTF-BP (Supplementary Fig. 15).

To explain this exceptionally high ionic conductivity, the $K^+$ transport pathway along the polymeric backbone of SCTF-BP was optimized and the two-dimensional free-energy landscape was generated based on the reported computational methodology (Fig. 3a and Supplementary Fig. 16)[23]. This shows the lowest energy barrier, with the critical steps involving the electrostatic interaction and dielectric interaction of $K^+$ ions with sulfonate groups and ether/triazine linkages, respectively. This is in good agreement with the simulation of mean square displacement for $K^+$ diffusion in SCTF membranes (Supplementary Fig. 17).

**Fig. 3: Ion transport across the SCTF-BP membrane.**




**a**, Modelling and calculation of free energy for the transport of $K^+$ across the SCTF-BP membrane matrix; the path 1-2-3-2-1 has the lowest free energy for $K^+$ transport. Insets demonstrate the specific interaction between $K^+$ and SCTF-BP chains in positions 1, 2 and 3. **b**, ssNMR measured for membrane samples of SCTF-BP, SPX-BP and Nafion. 100 mM NaCl in water was used as control, with membrane samples immersed in 0.1 M NaCl solution ($Na^+$ rather than $K^+$ because of improved NMR sensitivity). **c**, $^{23}$Na ssNMR for Nafion and SCTF-BP membranes in 0.1 M NaCl solution at varying temperature. **d**, PFG–NMR spectra collected for 0.1 M NaCl solution and

membrane samples of SPX-BP and SCTF-BP. Spectra for Nafion and SPPO are given in Supplementary Fig. 20. **e**,**f**, Plot of PFG–NMR signal intensity versus magnetic gradient strength (**e**, $B$ values) and diffusion coefficients derived from PFG–NMR for $Na^+$ in water, SCTF-BP, SPX-BP, Nafion and SPPO membrane samples (**f**). Echo profiles are fitted to the Stejskal–Tanner equation. $I$ denotes the echo height at a given gradient strength; $B$ denotes the product of all parameters before the diffusion coefficient ($D$) of the Stejskal–Tanner equation (that is, equation 9 in Supplementary Information); $r^1$ and $r^2$ denote the distance between the potassium ion and the geometric centre of two adjacent sulfonate groups. Error bars (s.d.) are derived from three measurements based on three individual membrane samples. **g**, Schematic showing electrostatic interaction and dielectric and steric effects during ion transport enabled by the negatively charged CTF membrane.

The computational work is supported by systematic NMR studies ($^{23}Na$ NMR was applied due to its higher sensitivity compared with K). $^{23}Na$ NMR can show the ion–dipole interaction between $Na^+$ ions and model triazine molecules in an aqueous solution (Supplementary Fig. 18). $^{23}Na$ magic-angle-spinning solid-state NMR ($^{23}Na$ ssNMR) shows two separate $Na^+$ signals for Nafion and SPX-BP membranes (Fig. 3b), with the downfield signal suggesting free $Na^+$ (located at the same position as that in 100 mM aqueous NaCl solution) and the upfield signal corresponding to $Na^+$ within membrane pores[19]. By contrast, only the downfield signal suggested that free $Na^+$ is detected for the SCTF-BP membrane (Fig. 3b). Notably, the full width at half maximum of the peak remains narrow and almost unchanged for SCTF-BP at 0–80 °C in the temperature-varied $^{23}Na$ ssNMR experiments, whereas an obvious 'motional' narrowing of the peak was observed for Nafion with increasing temperature (Fig. 3c and Supplementary Fig. 19). This line shape represents the temporally averaged local environment, which is closely related to the mobility of $Na^+$.

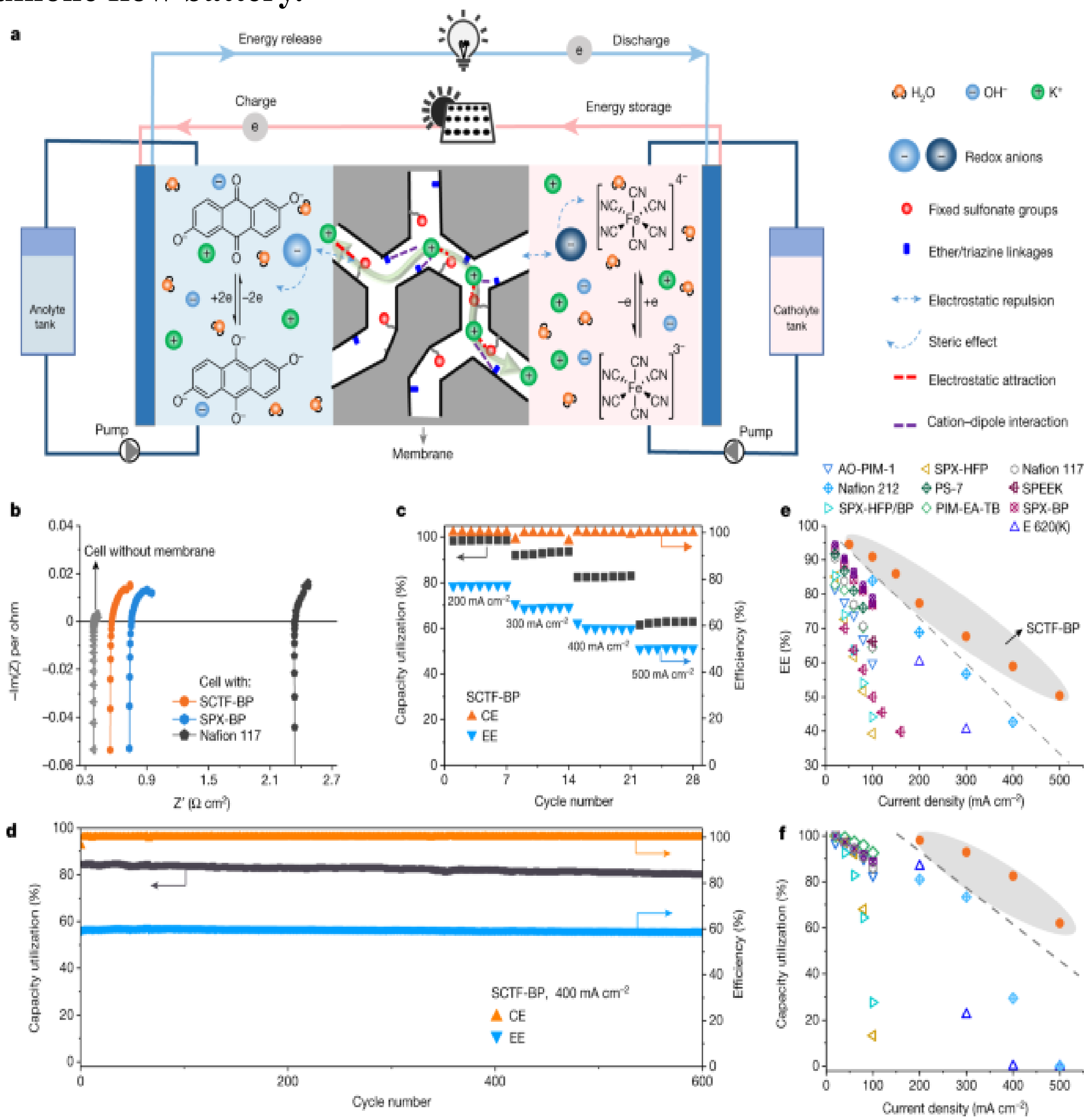
The mobility (diffusion coefficient) of $Na^+$ through these membranes was further measured by pulsed-field gradient–stimulated-echo nuclear magnetic resonance (PFG–NMR) experiments (Fig. 3d and Supplementary Fig. 20). By fitting echo profiles with the Stejskal–Tanner equation[28] (Fig. 3e) the

calculated $Na^+$ diffusion coefficient of SCTF-BP is $1.18 \times 10^{-9}$ $m^2$ $s^{-1}$, which is at least one order of magnitude greater than that of Nafion, SPPO and SPX-BP ($6.50 \times 10^{-11}$, $1.59 \times 10^{-10}$ and $2.86 \times 10^{-10}$ $m^2$ $s^{-1}$, respectively; Fig. 3f). Indeed, such high $Na^+$ mobility is close to the measured diffusion coefficient of $Na^+$ in water ($1.28 \times 10^{-9}$ $m^2$ $s^{-1}$; Fig. 3f), even at infinite dilution ($1.33 \times 10^{-9}$ $m^2$ $s^{-1}$), which can be considered as free diffusion and attests to the secondary interactions between cations and the pore wall surface and steric effect (Fig. 3g).

The advantages of CTF materials serving as selective ion-transport membranes were exemplified using aqueous organic redox flow batteries (AORFBs). For this application a membrane separator acts as both electrolyte insulator and charge carrier conductor between two half-cells (Fig. 4a). Current AORFBs can operate only at relatively low current densities, mostly at 100 mA $cm^{-2}$ or less, due to the high membrane resistance caused by the impeded diffusion of charge carrier ions traversing synthetic ion-exchange membranes. To assess the performance of SCTF-BP as a membrane separator in the alkaline quinone AORFB, full-flowing cells were assembled based on 2,6-DHAQ/$K_4Fe(CN)_6$ electrolytes and tested at both low (0.1 M) and high (0.4 M) electrolyte concentration (Fig. 4 and Supplementary Figs. 21 and 22). As controls, otherwise identical cells with commercial Nafion 117, Nafion 212 and Fumasep E-620(K) were assembled (Supplementary Figs. 23–26). The cell running at 0.1 M electrolyte concentration based on the SCTF-BP membrane showed a capacity utilization of over 85% at a current density of 120 mA $cm^{-2}$ (Supplementary Fig. 21); however, the cell using Nafion 117 could not even operate/cycle at the same current density, which shows capacity utilization decreasing sharply from 94.2 to 34.5% with current density increasing from 40 to 100 mA $cm^{-2}$ (Supplementary Fig. 23). The enhanced performance found with the SCTF-BP membrane is ascribed to its remarkably low resistance of 0.53 $\Omega$ $cm^2$, as shown by high-frequency EIS in cells operating at 50% state of charge (Supplementary Figs. 21 and 22). By further subtracting contributions from other components, the neat SCTF-BP membrane shows an areal-specific resistance as low as 0.17 $\Omega$ $cm^2$ (Fig. 4b) at an electrolyte concentration of 0.4 M and contributes only 20.7% of polarization resistance for the whole cell, which is much lower than that of Nafion 117, Nafion 212

and Fumasep E-620(K) in otherwise identical cells (Supplementary Table 4 and Supplementary Figs. 24–26).

**Fig. 4: SCTF-BP membrane enables rapid charging of aqueous alkaline quinone flow battery.**




**a**, Schematic illustrating an alkaline quinone flow battery assembled with the SCTF-BP membrane, and conduction of $K^+$ ions across the membrane matrix. The catholyte is ferrocyanide and the anolyte molecule is DHAQ. **b**, EIS spectra measured in cells assembled with SCTF-BP, SPX-BP and Nafion 117 membrane, respectively. The grey line represents the EIS

spectrum of the cell without a membrane. **c**, Coulombic efficiency (CE), capacity utilization and EE of the cell assembled with SCTF-BP at varying current density. For each current density, seven repetitions were performed to ensure accuracy. Fluctuations were observed only when current density was switched. **d**, Galvanostatic cycling at a current density of 400 mA $cm^{-2}$ for cells assembled with a SCTF-BP membrane. **b**–**d**, Electrolyte concentration 0.4 M. **e**,**f**, Energy efficiency (**e**) and capacity utilization (**f**) of quinone flow batteries assembled with commercial, PIM and SCTF-BP membranes are plotted as functions of current density. Symbols shown in inset in **e** apply also to **f**. **e**,**f**, Dashed lines and shading are a visual guide only; detailed values provided in Supplementary Table 5.

Such a low resistance of the SCTF-BP membrane enables cell operation at extreme current densities, from 200 to 500 mA $cm^{-2}$, but also shows exceptionally high energy efficiency (round-trip) and high-capacity utilization at 0.4 M electrolyte concentration (Fig. 4c–f). For example, this cell achieved capacity utilization of 62% and an energy efficiency (EE) of 50.4% at current density up to 500 mA $cm^{-2}$ (Fig. 4c), approaching the levels of acidic quinone AORFBs in which the charge carrier ($H^+$) diffuses much more rapidly than $K^+$ (ref. [29]), confirming the potential of SCTF-BP membranes for rapid charging of alkaline quinone AORFBs. Furthermore, flow batteries based on the SCTF-BP membrane achieved long-term galvanostatic cell operation over 5,000 cycles at 160 mA $cm^{-2}$ using 0.1 M electrolyte concentration (Supplementary Fig. 21), and over 600 cycles at 400 mA $cm^{-2}$ using 0.4 M electrolyte concentration (Fig. 4d). In addition to low membrane resistance, the SCTF-BP membrane can effectively block the cross-contamination of active species during cell cycling at both room temperature and 45 °C (Supplementary Fig. 27 and Supplementary Table 6). The exceptional performance of alkaline AORFBs encouraged us to extend the application of CTF membranes to pH-neutral AORFBs using nitroxide radical/viologen-based redox chemistry (Supplementary Figs. 28 and 29). It is clear that our CTF membranes enable both alkaline and pH-neutral AORFBs to perform rapid charging at extreme current density with concomitantly high EE and high-capacity utilization. These data far exceed those for otherwise identical cells assembled with commercial membranes and state-of-the-art ion-sieving membranes (Supplementary Tables 5, 7 and 8).

To conclude, microporous CTF membranes—derived from aromatic nitrile monomers with different charge functionalities through a superacid-catalysed organic sol-gel process—achieved high-performing AORFBs for a range of redox chemistries, attributable to the low energy barriers for cation diffusion through well-confined micropores. This design strategy may be broadly applicable considering the various options of organic reactions and monomers in construction of covalent organic frameworks. We anticipate that these CTF membranes will enable a broader range of applications beyond those presently demonstrated in AORFBs—for instance, in electrochemical devices for $CO_2$ reduction[10]. Although a local frictionless ion flow related to the wettability of 3D nanochannels[30] was observed within the SCTF-BP membrane, further manipulations of CTF membrane architectures are required to facilitate long-range ion diffusion, which may be hindered by pore tortuosity. The concept we propose here extends the potential of ultrafast ion-conducting membranes beyond inorganic microporous counterparts (for example, zeolites, carbon nanotubes, graphene and MXenes) and suggests opportunities for the development of polymer membranes relying on secondary interactions for precise molecular separation.

## Data availability

The data that support the findings of this study are available in the manuscript and its Supplementary Information, and also from the corresponding authors on reasonable request.

## References


1. Salvatore, D. A. et al. Designing anion exchange membranes for $CO_2$ electrolysers. *Nat. Energy* **6**, 339–348 (2021).

2. Lagadec, M. F. & Grimaud, A. Water electrolysers with closed and open electrochemical systems. *Nat. Mater.* **19**, 1140–1150 (2020).

3. Jiao, K. et al. Designing the next generation of proton-exchange membrane fuel cells. *Nature* **595**, 361–369 (2021).

4. Tang, H. et al. Fuel cells with an operational range of –20 °C to 200 °C enabled by phosphoric acid-doped intrinsically ultramicroporous membranes. *Nat. Energy* **7**, 153–162 (2022).

5. Xiong, P., Zhang, L., Chen, Y., Peng, S. & Yu, G. A chemistry and microstructure perspective on ion-conducting membranes for redox flow batteries. *Angew. Chem. Int. Ed. Engl.* **60**, 24770–24798 (2021).

6. Zhang, L., Feng, R., Wang, W. & Yu, G. Emerging chemistries and molecular designs for flow batteries. *Nat. Rev. Chem.* **6**, 524–543 (2022).

7. Uliana, A. A. et al. Ion-capture electrodialysis using multifunctional adsorptive membranes. *Science* **372**, 296–299 (2021).

8. Epsztein, R., DuChanois, R. M., Ritt, C. L., Noy, A. & Elimelech, M. Towards single-species selectivity of membranes with subnanometre pores. *Nat. Nanotechnol.* **15**, 426–436 (2020).

9. Zhou, X. et al. Intrapore energy barriers govern ion transport and selectivity of desalination membranes. *Sci. Adv.* **6**, eabd9045 (2020).

10. Lees, E. W., Mowbray, B. A. W., Parlane, F. G. L. & Berlinguette, C. P. Gas diffusion electrodes and membranes for $CO_2$ reduction electrolysers. *Nat. Rev. Mater.* **7**, 55–64 (2021).

11. Ran, J. et al. Ion exchange membranes: new developments and applications. *J. Membr. Sci.* **522**, 267–291 (2017).

12. Foglia, F. et al. Disentangling water, ion and polymer dynamics in an anion exchange membrane. *Nat. Mater.* **21**, 555–563 (2022).

13. Shin, D. W., Guiver, M. D. & Lee, Y. M. Hydrocarbon-based polymer electrolyte membranes: importance of morphology on ion transport and membrane stability. *Chem. Rev.* **117**, 4759–4805 (2017).

14. Kusoglu, A. & Weber, A. Z. New insights into perfluorinated sulfonic-acid ionomers. *Chem. Rev.* **117**, 987–1104 (2017).

15. Park, H. B., Kamcev, J., Robeson, L. M., Elimelech, M. & Freeman, B. D. Maximizing the right stuff: the trade-off between membrane permeability and selectivity. *Science* **356**, eaab0530 (2017).

16. Thompson, K. A. et al. N-Aryl–linked spirocyclic polymers for membrane separations of complex hydrocarbon mixtures. *Science* **369**, 310–315 (2020).

17. Lai, H. W. H. et al. Hydrocarbon ladder polymers with ultrahigh permselectivity for membrane gas separations. *Science* **375**, 1390–1392 (2022).

18. Baran, M. J. et al. Design rules for membranes from polymers of intrinsic microporosity for crossover-free aqueous electrochemical devices. *Joule* **3**, 2968–2985 (2019).

19. Tan, R. et al. Hydrophilic microporous membranes for selective ion separation and flow-battery energy storage. *Nat. Mater.* **19**, 195–202 (2020).

20. Zuo, P. et al. Sulfonated microporous polymer membranes with fast and selective ion transport for electrochemical energy conversion and storage. *Angew. Chem. Int. Ed. Engl.* **59**, 9564–9573 (2020).

21. Ye, C. et al. Development of efficient aqueous organic redox flow batteries using ion-sieving sulfonated polymer membranes. *Nat. Commun.* **13**, 3184 (2022).

22. Ye, C. et al. Long-life aqueous organic redox flow batteries enabled by amidoxime-functionalized ion-selective polymer membranes. *Angew. Chem. Int. Ed. Engl.* **61**, e202207580 (2022).

23. Baran, M. J. et al. Diversity-oriented synthesis of polymer membranes with ion solvation cages. *Nature* **592**, 225–231 (2021).

24. Low, Z. X., Budd, P. M., McKeown, N. B. & Patterson, D. A. Gas permeation properties, physical aging, and its mitigation in high free volume glassy polymers. *Chem. Rev.* **118**, 5871–5911 (2018).

25. Koros, W. J. & Zhang, C. Materials for next-generation molecularly selective synthetic membranes. *Nat. Mater.* **16**, 289–297 (2017).

26. Zhu, X. et al. A superacid-catalyzed synthesis of porous membranes based on triazine frameworks for $CO_2$ separation. *J. Am. Chem. Soc.* **134**, 10478–10484 (2012).

27. Yang, Z. et al. Surpassing Robeson upper limit for $CO_2/N_2$ separation with fluorinated carbon molecular sieve membranes. *Chem* **6**, 631–645 (2020).

28. Johnson, C. S. Jr Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications. *Prog. Nucl. Magn. Reson. Spectrosc.* **34**, 203–256 (1999).

29. Huskinson, B. et al. A metal-free organic-inorganic aqueous flow battery. *Nature* **505**, 195–198 (2014).

30. Zhang, X., Liu, H. & Jiang, L. Wettability and applications of nanochannels. *Adv. Mater.* **31**, 1804508 (2019).

## Acknowledgements

This work was financially supported by the National Key Research and Development Project (nos. 2021YFB4000300 and 2020YFB1505600) and the National Natural Science Foundation of China (nos. 21922510, 21878281, U20A20127 and 52021002). T.L.L. and J.L. acknowledge financial support by the National Science Foundation (career award, grant no. 1847674). N.B.M. and C.Y. thank the Engineering and Physical Science Research Council for funding through programme grant SynHiSel (no. EP/V047078). P.Z. acknowledges financial support by the China Postdoctoral Science Foundation (no. 2021M693066) and Fundamental Research Funds for the Central Universities (no. WK2060000028). The authors thank H. Zhang for help with positron annihilation lifetime spectroscopy experiments. Z.Y. thanks the support of his wife, Q. Fang, and his sons, G.-C. Yang and G.-Q. Yang, during the difficult time of the COVID-19 pandemic.

# Author information

Author notes

1. These authors contributed equally: Peipei Zuo, Chunchun Ye, Zhongren Jiao


## Authors and Affiliations

1. Key Laboratory of Precision and Intelligent Chemistry, Department of Applied Chemistry, School of Chemistry and Material Science, University of Science and Technology of China, Hefei, P. R. China

   Peipei Zuo, Zhongren Jiao, Junkai Fang, Zhengjin Yang & Tongwen Xu

2. EastCHEM School of Chemistry, University of Edinburgh, Edinburgh, UK

   Chunchun Ye & Neil B. McKeown

3. Utah State University, Chemistry and Biochemistry, Logan, UT, USA

   Jian Luo & T. Leo Liu

4. Laboratory of Organic and Macromolecular Chemistry, Friedrich Schiller University Jena, Jena, Germany

   Ulrich S. Schubert

5. Center for Energy and Environmental Chemistry Jena, Friedrich Schiller University Jena, Jena, Germany

   Ulrich S. Schubert


## Contributions

Z.Y. and T.X. conceived and directed the project. T.L.L. provided suggestions and guidance during project implementation. P.Z., C.Y. and J.F.

conducted experiments. Z.J. performed molecular simulations. P.Z., C.Y., Z.J., J.L., T.L.L., Z.Y. and T.X. analysed and discussed data. P.Z., C.Y., Z.J., Z.Y. and T.X. wrote the manuscript with input from T.L.L., J.L., J.F., U.S.S. and N.B.M. All authors offered constructive feedback on the manuscript.


### Corresponding authors

Correspondence to T. Leo Liu, Zhengjin Yang or Tongwen Xu.


## Ethics declarations

### Competing interests

The authors declare no competing interests.

## Peer review

### Peer review information

*Nature* thanks Young Moo Lee and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

## Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Supplementary information

### Supplementary Information

This file contains Supplementary Methods, Figs. 1–29, Tables 1–8 and References.

# Rights and permissions

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Reprints and Permissions

# About this article

## Cite this article

Zuo, P., Ye, C., Jiao, Z. *et al.* Near-frictionless ion transport within triazine framework membranes. *Nature* **617**, 299–305 (2023). https://doi.org/10.1038/s41586-023-05888-x

- Received: 02 August 2022
- Accepted: 24 February 2023
- Published: 26 April 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05888-x

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05888-x

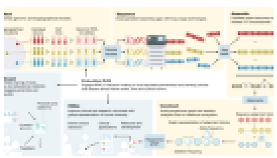
| Section menu | Main menu |



## Towards a reference genome that captures global genetic diversity

30 October 2020

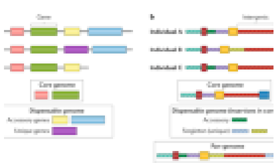
Karen H. Y. Wong, Walfred Ma, … Pui-Yan Kwok



## The Human Pangenome Project: a global resource to map genomic diversity

20 April 2022

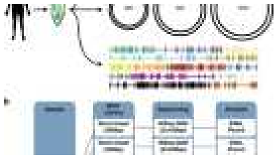
Ting Wang, Lucinda Antonacci-Fulton, … the Human Pangenome Reference Consortium



## Pan-genomics in the human genome era

07 February 2020

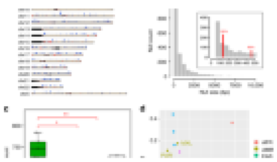
Rachel M. Sherman & Steven L. Salzberg



## [Extensive and deep sequencing of the Venter/HuRef genome for developing and benchmarking genome analysis tools]

18 December 2018

Bo Zhou, Joseph G. Arthur, … Alexander E. Urban



## [De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations]

02 August 2018

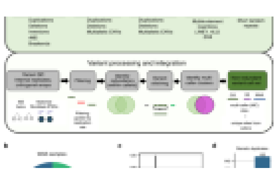
Karen H. Y. Wong, Michal Levy-Sakin & Pui-Yan Kwok



## [Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes]

11 April 2022

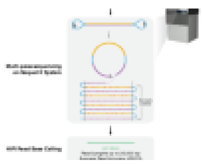
Jana Ebler, Peter Ebert, … Tobias Marschall



[**Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats**]

10 June 2020

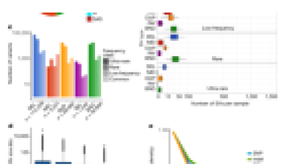
David Jakubosky, Erin N. Smith, … Kelly A. Frazer



[**Highly accurate long-read HiFi sequencing data for five complex genomes**]

17 November 2020

Ting Hon, Kristin Mars, … David R. Rank



[**Mapping and characterization of structural variation in 17,795 human genomes**]

27 May 2020

Haley J. Abel, David E. Larson, … Ira M. Hall

- Article
- Open Access
- Published: 10 May 2023

# A draft human pangenome reference


- Wen-Wei Liao,
- Mobin Asri,
- Jana Ebler,
- Daniel Doerr,
- Marina Haukness,
- Glenn Hickey,
- Shuangjia Lu,
- Julian K. Lucas,
- Jean Monlong,
- Haley J. Abel,
- Silvia Buonaiuto,
- Xian H. Chang,
- Haoyu Cheng,
- Justin Chu,
- Vincenza Colonna,
- Jordan M. Eizenga,
- Xiaowen Feng,
- Christian Fischer,
- Robert S. Fulton,
- Shilpa Garg,
- Cristian Groza,
- Andrea Guarracino,
- William T. Harvey,
- Simon Heumos,
- Kerstin Howe,
- Miten Jain,
- Tsung-Yu Lu,
- Charles Markello,
- Fergal J. Martin,
- Matthew W. Mitchell,
- Katherine M. Munson,
- Moses Njagi Mwaniki,

- Adam M. Novak,
- Hugh E. Olsen,
- Trevor Pesout,
- David Porubsky,
- Pjotr Prins,
- Jonas A. Sibbesen,
- Jouni Sirén,
- Chad Tomlinson,
- Flavia Villani,
- Mitchell R. Vollger,
- Lucinda L. Antonacci-Fulton,
- Gunjan Baid,
- Carl A. Baker,
- Anastasiya Belyaeva,
- Konstantinos Billis,
- Andrew Carroll,
- Pi-Chuan Chang,
- Sarah Cody,
- Daniel E. Cook,
- Robert M. Cook-Deegan,
- Omar E. Cornejo,
- Mark Diekhans,
- Peter Ebert,
- Susan Fairley,
- Olivier Fedrigo,
- Adam L. Felsenfeld,
- Giulio Formenti,
- Adam Frankish,
- Yan Gao,
- Nanibaa' A. Garrison,
- Carlos Garcia Giron,
- Richard E. Green,
- Leanne Haggerty,
- Kendra Hoekzema,
- Thibaut Hourlier,
- Hanlee P. Ji,
- Eimear E. Kenny,

- Barbara A. Koenig,
- Alexey Kolesnikov,
- Jan O. Korbel,
- Jennifer Kordosky,
- Sergey Koren,
- HoJoon Lee,
- Alexandra P. Lewis,
- Hugo Magalhães,
- Santiago Marco-Sola,
- Pierre Marijon,
- Ann McCartney,
- Jennifer McDaniel,
- Jacquelyn Mountcastle,
- Maria Nattestad,
- Sergey Nurk,
- Nathan D. Olson,
- Alice B. Popejoy,
- Daniela Puiu,
- Mikko Rautiainen,
- Allison A. Regier,
- Arang Rhie,
- Samuel Sacco,
- Ashley D. Sanders,
- Valerie A. Schneider,
- Baergen I. Schultz,
- Kishwar Shafin,
- Michael W. Smith,
- Heidi J. Sofia,
- Ahmad N. Abou Tayoun,
- Françoise Thibaud-Nissen,
- Francesca Floriana Tricomi,
- Justin Wagner,
- Brian Walenz,
- Jonathan M. D. Wood,
- Aleksey V. Zimin,
- Guillaume Bourque,
- Mark J. P. Chaisson,

- Paul Flicek,
- Adam M. Phillippy,
- Justin M. Zook,
- Evan E. Eichler,
- David Haussler,
- Ting Wang,
- Erich D. Jarvis,
- Karen H. Miga,
- Erik Garrison ,
- Tobias Marschall ,
- Ira M. Hall ,
- Heng Li &
- …
- Benedict Paten

Show authors

*Nature* **volume 617**, pages 312–324 (2023)

- 2061 Altmetric

- Metrics details

## Abstract

Here the Human Pangenome Reference Consortium presents a first draft of the human pangenome reference. The pangenome contains 47 phased, diploid assemblies from a cohort of genetically diverse individuals[1]. These assemblies cover more than 99% of the expected sequence in each genome and are more than 99% accurate at the structural and base pair levels. Based on alignments of the assemblies, we generate a draft pangenome that captures known variants and haplotypes and reveals new alleles at structurally complex loci. We also add 119 million base pairs of euchromatic polymorphic sequences and 1,115 gene duplications relative to the existing reference GRCh38. Roughly 90 million of the additional base pairs are derived from structural variation. Using our draft pangenome to analyse short-read data reduced small variant discovery errors by 34% and increased

the number of structural variants detected per haplotype by 104% compared with GRCh38-based workflows, which enabled the typing of the vast majority of structural variant alleles per sample.

# Main

The human reference genome has formed the backbone of human genomics since its initial draft release more than 20 years ago[2]. The primary sequences are a mosaic representation of individual haplotypes containing one representative scaffold sequence for each chromosome. There are 210 Mb of gap or unknown (151 Mb) or computationally simulated sequences (59 Mb) within the current GRCh38 release, constituting 6.7% of the primary chromosome scaffolds. Missing reference sequences create an observational bias, or streetlamp effect, which limits studies to be within the boundaries of the reference. Recently, the Telomere-to-Telomere (T2T) consortium finished the first complete sequence of a haploid human genome, T2T-CHM13, which provides a contiguous representation of each autosome and of chromosome X, with the exception of some ribosomal DNA arrays that remain to be fully resolved[3]. Using T2T-CHM13 directly improves genomic analyses; for example, discovering 3.7 million additional single-nucleotide polymorphisms (SNPs) in regions non-syntenic to GRCh38 and better representing the true copy number variants (CNVs) of samples from the 1000 Genomes Project (1KG) compared with GRCh38 (refs. [1,4]).

Although T2T-CHM13 represents a major achievement, no single genome can represent the genetic diversity of our species. Previous studies have identified tens of megabases of sequence contained within structural variants (SVs) that are polymorphic within the population[5]. Owing to the absence of these alternative alleles from the reference genome, more than two-thirds of SVs have been missed in studies that used short-read data and the human reference assembly[6,7,8], despite individual SVs being more likely to affect gene function than either individual SNPs or short insertions and deletions (indels)[9,10].

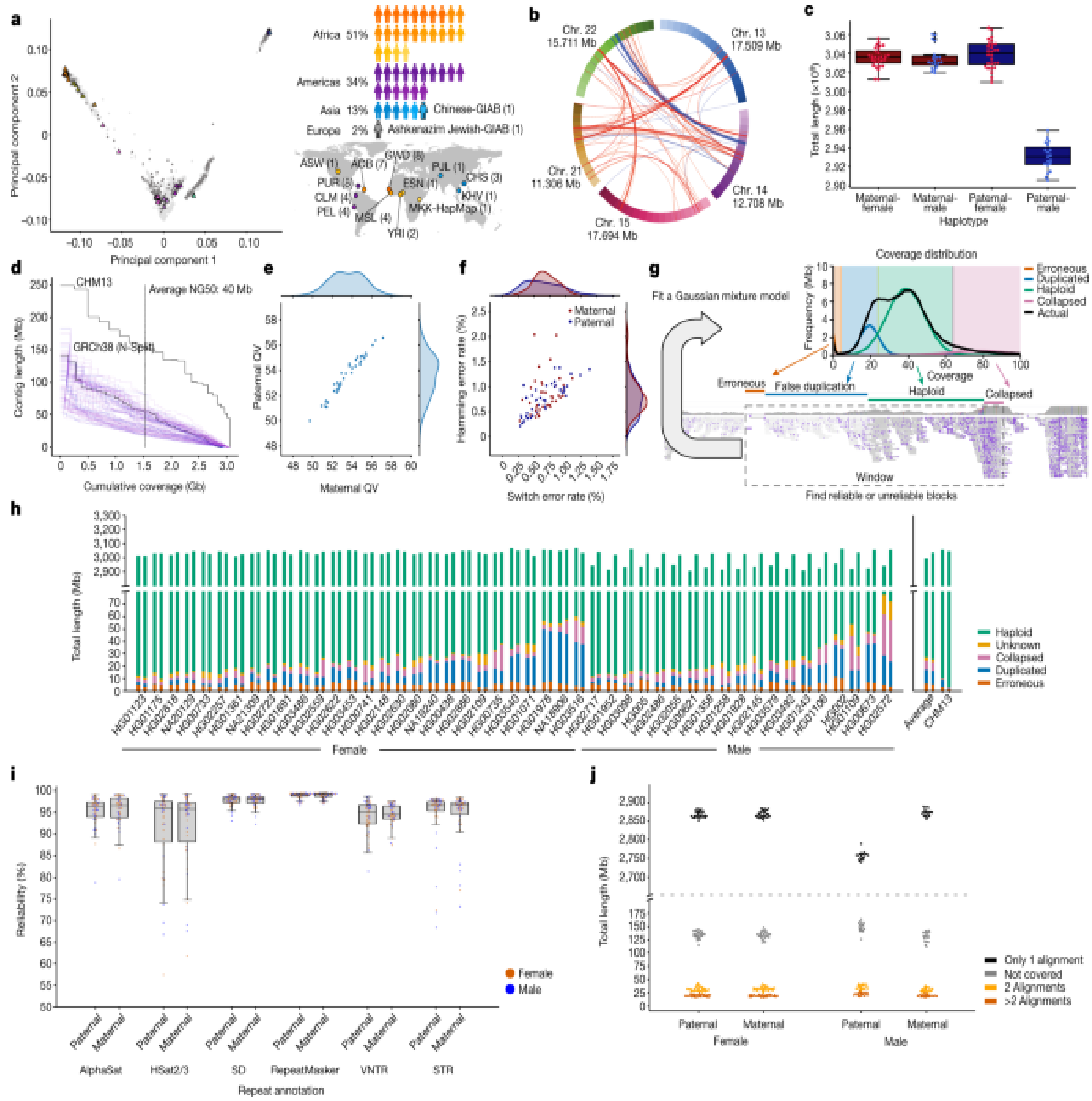
To overcome reference bias, a transition to a pangenomic reference has been envisioned[11,12]. Pangenomic methods have rapidly progressed over the past

few years[13,14,15] such that it is now practical to propose that common genomic analyses use a pangenome. Here we sequence and assemble a set of diverse individual genomes and present a draft human pangenome, the first release from the Human Pangenome Reference Consortium (HPRC)[15]. These genomes represent a subset of the planned HPRC panel, which aims to better capture global genomic diversity across the 700 haplotypes of 350 individuals.

## Assembling 47 diverse human genomes

We assembled 47 fully phased diploid assemblies from genomes selected to represent global genetic diversity (Fig. 1a) and for which consent had been given for unrestricted access. All assemblies have been made publicly available, along with all data and analyses. These assemblies include 29 samples with long and linked read sequencing data generated entirely by the HPRC and 18 samples sequenced by other efforts[16,17,18]. In some cases, we supplemented the 18 additional samples with further sequencing. We selected the 29 HPRC samples from the 1KG lymphoblastoid cell lines, limiting selection to those lines classified as karyotypically normal and with low passage (to avoid artefacts from cell culture). We also ensured that the cell lines were derived from participants for whom whole-genome sequencing (WGS) data were available for both parents (for haplotype phasing). Cell lines meeting these criteria were prioritized by genetic and biogeographic diversity (Methods).

**Fig. 1: Presenting 47 accurate and near-complete diverse diploid human genome assemblies.**



**a**, Selecting the HPRC samples. Left, the first two principal components of 1KG samples showing HPRC (triangles) samples, excluding HG002, HG005 and NA21309. Right, summary of the HPRC sample subpopulations (three letter abbreviations) on a map of Earth as defined by the 1KG. ACB, African Caribbean in Barbados; ASW, African Ancestry in Southwest US; CHS, Han Chinese South; CLM, Colombian in Medellin, Colombia; ESN, Esan in Nigeria; GWD, Gambian in Western Division; KHV, Kinh in Ho Chi Minh City, Vietnam; MKK, Maasai in Kinyawa, Kenya; MSL, Mende in Sierra Leone; PEL, Peruvian in Lima, Peru; PJL, Punjabi in Lahore, Pakistan; PUR, Puerto Rican in Puerto Rico; YRI, Yoruba in Ibadan, Nigeria. **b**,

Interchromosomal joins between acrocentric chromosome short arms. Red, the join is on the same strand; blue, otherwise. **c**, Total assembled sequence per haploid phased assembly. **d**, Assembly contiguity shown as a NGx plot. T2T-CHM13 and GRCh38 contigs are included for comparison. **e**, Assembly QVs showing the base-level accuracy of the maternal and paternal assembly for each sample. **f**, Yak-reported phasing accuracy showing the switch error percentage versus Hamming error percentage. **g**, Flagger read-based assembly evaluation pipeline. Coverage is calculated across the genome and a mixture model is fit to account for reliably assembled haploid sequence and various classes of unreliably assembled sequence. For each coverage block, a label is assigned according to the most probable mixture component to which it belongs: erroneous, falsely duplicated, (reliable) haploid, collapsed, and unknown. **h**, Reliability of the 47 HPRC assemblies using read mapping. For each sample, the left bar is the paternal and the right bar is the maternal haplotype. Regions flagged as haploid are reliable (green), constituting more than 99% on average of each assembly. The *y* axis is broken to show the dominance of the reliable haploid component and the stratification of the unreliable blocks. **i**, Assembly reliability of six types of repeats. AlphaSat, alpha satellites; HSat2/3, human satellites 2 and 3. **j**, Completeness of the HPRC assemblies relative to T2T-CHM13. The number of reference bases covered by none, by one, by two or by more than two alignments are included.

We created a consistent set of deeply sequenced data types for every sample (Supplementary Table 1). The data included Pacific Biosciences (PacBio) high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) long-read sequencing, Bionano optical maps and high-coverage Hi-C Illumina short-read sequencing for all HPRC samples. We also gathered previously generated high-coverage Illumina sequencing data for both parents of each participant[19]. We generated on average 39.7× HiFi sequence depth of coverage for the 46 HPRC samples (excluding HG002, which had around 130× coverage). This depth of coverage is consistent with the requirements for high-quality, state-of-the-art assemblies[20] and facilitates comprehensive variant discovery irrespective of allele frequency (AF). The N50 value, which represents the shortest read length at which 50% of the total sequenced bases are covered by considering only equal or longer reads, was 19.6 kb on average for the HiFi reads (Supplementary Table 1; excluding

HG002 because it was sequenced using a different library preparation protocol).

For the core assembler, we chose Trio-Hifiasm[20] after detailed benchmarking of several alternatives[21]. Trio-Hifiasm uses PacBio HiFi long-read sequences and parental Illumina short-read sequences to produce near fully phased contig assemblies. The complete assembly pipeline (Supplementary Fig. 1 and Methods) included steps to remove adaptor and nonhuman sequence contamination and to ensure a single mitochondrial assembly per maternal assembly.

## Assembly assessment

We first searched for large-scale misassemblies, looking for gene duplication errors, phasing errors and interchromosomal misjoins (Methods). We manually fixed three large duplication errors and one large phasing error, but left smaller errors, which are difficult to definitively distinguish from SVs. We found 217 putative interchromosomal joins. Only one of these joins (in the paternal assembly of HG02080) was located in a euchromatic, non-acrocentric region and was manually confirmed to be a misassembly. The remaining joins involved the short arms of the acrocentric chromosomes (Fig. 1b and Supplementary Table 2). This may be the result of misalignment, nonallelic gene conversion or other biological mechanisms that maintain large-scale homology between the short arms of the acrocentrics—a phenomenon that we have studied in an associated paper[22].

To evaluate the resulting assemblies after manual correction of errors, we developed an automated assembly quality control pipeline that combined methods to assess the completeness, contiguity, base level quality and phasing accuracy of each assembly (Supplementary Table 3 and Methods). Haploid assemblies containing an X chromosome averaged a total length of 3.04 Gb, 99.3% the length of T2T-CHM13 (3.06 Gb), which also contains an X chromosome. Haploid assemblies containing a Y chromosome averaged a total length of 2.93 Gb, which reflects the size difference between the sex chromosomes (Fig. 1c). The average NG50 value, a widely used measure of contiguity, was comparable with the contig NG50 of GRCh38 (40 Mb compared with 56 Mb, respectively; Fig. 1d). Using short substrings (*k*-mer

values of 31) derived from Illumina data, the Yak *k*-mer analyser[20] estimated an average quality value (QV) of 53.57 for the assemblies, which corresponded to an average of 1 base error per 227,509 bases (Fig. 1e). To validate these QV estimates, we benchmarked the HG002 and HG005 assembly-based variant calls against the small variants called using Genome in a Bottle (GIAB; v.4.2.1). We estimated QVs of 54 for HG002 and 55 for HG005, which were highly similar to the *k*-mer QVs estimated using Yak. Consistent with our manual observation that most errors were primarily small indels in low-complexity regions, we found that approximately 32% of the indel errors were in homopolymers longer than 5 bp and an additional 48% were in tandem repeats and low-complexity regions. Moreover, about 42% of the indel errors were genotype errors, mostly heterozygous variants incorrectly called as homozygous variants due to collapsed haplotypes in the two assemblies of an individual (Supplementary Table 4). We next used Yak to analyse the phasing accuracy between the maternal and paternal assemblies using *k*-mer values derived from Illumina sequencing of the parents. An average haplotype switch error rate of 0.67% was observed and a Hamming error rate of 0.79% (Fig. 1f). We also calculated phase accuracy using Pstools[23,24], which uses Hi-C sequence data of the sample not used to create the assembly. Pstools reported slightly lower switch error rates than Yak but comparable Hamming error rates (Supplementary Fig. 2). Taken together, the above results indicate that the assemblies are highly contiguous and accurate.

## Regional assembly reliability

To determine which portions of the assemblies are reliable, we developed a read-based pipeline, Flagger, that detects different types of misassemblies within a phased diploid assembly (Fig. 1g and Methods). The pipeline works by mapping the HiFi reads to the combined maternal and paternal assembly in a haplotype-aware manner. It then identifies coverage inconsistencies within these read mappings that are likely to be due to assembly errors. This process is similar to likelihood-based approaches, which assess the assembly given the reads[25], but is adapted to work with long reads and diploid assemblies. Using Flagger, we identified only 0.88% (26.4 Mb) of each assembly as unreliable (Fig. 1h and Supplementary Table 5). Using T2T-CHM13, we estimated that 0.09% of reliably assembled blocks were falsely

labelled as unreliable (Methods). Compared with the distribution of contig sizes, the unreliable blocks were short (54.6 kb N50 average). We intersected the unreliable blocks in the assemblies from Flagger with different repeat annotations (Fig. 1i and Supplementary Table 6). We estimated that the following percentage of elements were correctly assembled: 95.4% of alpha satellites; 91.5% of human satellites 2 and 3; 97.7% of segmental duplications (SDs); 94.3% of variable number tandem repeats (VNTRs); 94.2% of short tandem repeats (STRs); and 98.8% of all human repeats[26].

## Completeness and CNV

To assess the completeness and copy number polymorphism of the assemblies, we aligned them to T2T-CHM13 (Methods). The paternal assemblies of male samples covered about 92.8% of T2T-CHM13 (excluding chromosome X) on average with exactly one alignment. For all other assemblies (excluding chromosome Y), about 94.1% on average was single-copy covered (Fig. 1j and Supplementary Table 7). On average, around 136 Mb (4.4%) of T2T-CHM13 was not covered by any alignment, which indicates that some parts of the genome are either systematically unassembled or cannot be reliably aligned. About 90% of these regions were centromeric or pericentromeric[27] (Extended Data Fig. 1). Despite the majority of unaligned bases occurring within and around centromeres, on average, 90% of divergent and monomeric alpha satellites, gamma satellites and centromeric transition regions were covered by at least one alignment. Excluding the T2T-CHM13 centromere and satellites[3] and including only the expected sex chromosome for each haploid assembly, on average, around 99.12% of the remaining reference was covered by exactly one alignment (Supplementary Table 7).

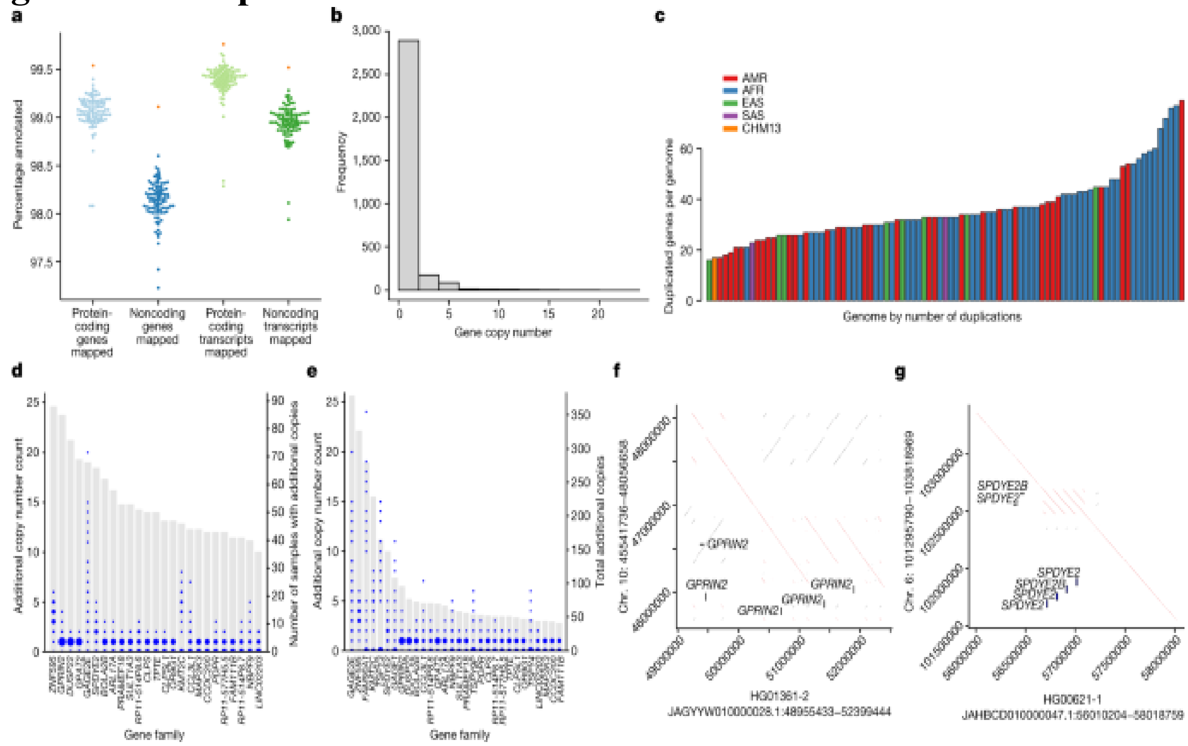
The average number of T2T-CHM13 bases with two or with more than two alignments was about 32.4 Mb (around 1.0%) and about 20.0 Mb (around 0.6%), respectively. On average, per haploid assembly, these duplicated regions had about 82.20% and 39.82% overlap with the pericentomeric or centromeric satellites and SDs, respectively, and around 94.62% had overlap with either of them. We characterized the accuracy of regions aligned to SDs in T2T-CHM13 (excluding chromosome Y) using a liftover of the assembly read-depth-based evaluation (Extended Data Fig. 2). On average, we

estimated that only 2.5% (4.99 out of 199 Mb) of the SD sequence that could be lifted onto T2T-CHM13 was in error according to the read depth. To identify SDs associated with these errors, we took all 5 kb windows across the unreliable regions and intersected them with the longest and most identical overlapping SD. The median length of SDs overlapping sequences in error was 3.0 times longer (288 kb compared with 96.3 kb) than those in correctly assembled SDs and 1.8% more identical (98.9 compared with 97.1). This result reinforces earlier findings that the length and identity of SDs play an important part in assembly accuracy[28].

## Annotating 47 diverse genomes

We developed a new Ensembl mapping pipeline to annotate GENCODE[29] genes and transcripts within each new haploid assembly (Methods). A median of 99.07% of protein-coding genes (range = 98.08–99.40%) and 99.42% of protein-coding transcripts (range = 98.29–99.66%) were identified in each of the HPRC assemblies (Fig. 2a and Supplementary Table 8). Similarly, a median of 98.16% of noncoding genes (range = 97.23–98.60%) and 98.96% of noncoding transcripts (range = 97.94–99.28%) were similarly annotated. Running this pipeline on T2T-CHM13 produced similar, slightly higher, results. Intersecting the HPRC annotations with the assembly reliability predictions, a median of 99.53% of gene and 99.79% of transcript annotations occurred wholly within reliable regions, which indicated that most of the annotated transcript haplotypes were structurally correct. To examine transcriptome base accuracy, we looked for nonsense and frameshift mutations in the set of canonical transcripts (one representative transcript per gene; Supplementary Fig. 3, Supplementary Table 8 and Methods). We found a median of 25 nonsense mutations and 72 frameshifts per assembly. A median of 21 (84%) and 58 (80%) of these nonsense mutations and frameshifts per assembly, respectively, were supported by the independently generated Illumina variant call sets. These numbers were within the range of previously reported numbers of loss-of-function mutations (between 10 and 150 per person, depending on the level of conservation of the mutation)[1,30]. Conservatively, if all the non-confirmed frameshifts and nonsense mutations are assembly errors, this would predict 18 such transcript-altering errors per transcriptome (1 per 1.7 million assembled transcriptome bases).

**Fig. 2: Transcriptome annotation of the assemblies.**




**a**, Ensembl mapping pipeline results. Percentages of protein-coding and noncoding genes and transcripts annotated from the reference set in each of the HPRC assemblies. Orange points represent T2T-CHM13 for comparison. **b**, Frequency of gene copy number. Individual genes may have separate copy number states among genomes, and the frequency reflects 3,210 observed copy number changes among the HPRC genomes. **c**, Number of distinct duplicated genes or gene families per phased assembly relative to the number of duplicated genes annotated in GRCh38 ($n = 152$). The GRCh38 gene duplications reflect families of duplicated genes, whereas the counts in other genomes reflect gene duplication polymorphisms. The assemblies are colour coded according to their population of origin. **d**, The top 25 most commonly CNV genes or gene families in the HPRC assemblies out of all 1,115 duplicated genes, ordered by the number of samples with additional copies relative to GRCh38. Grey bars, the number of samples with additional copies. Blue circles, the number of additional copies per sample, with the size of the circle proportional to the number of samples. **e**, The top 30 most individually copied CNV genes or gene families in the HPRC assemblies, ordered by total number of additional copies observed. Blue circles, the number of additional copies per sample. Grey bars, the total

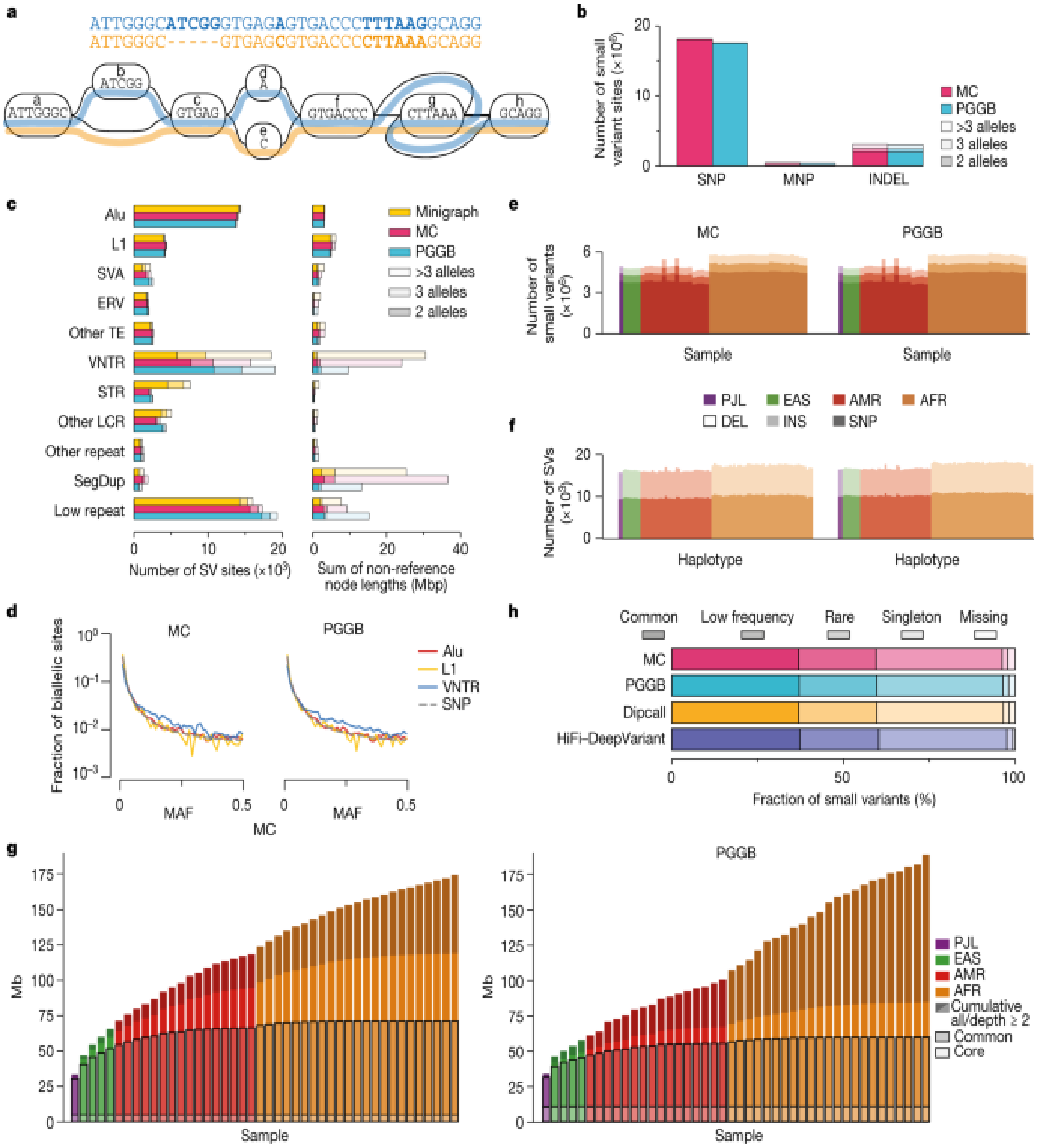
number of additional copies summed over the samples. **f**, Dotplot illustrating haplotype-resolved *GPRIN2* gains in the HG01361 assembly relative to GRCh38. **g**, Dotplot illustrating *SPDYE2–SPDYE2B* haplotype resolved gains within a tandem duplication cluster of the HG00621 assembly relative to GRCh38.

There were 1,115 protein-coding gene families within the Flagger-predicted reliable regions of the full set of assemblies that had a gain in copy number in at least one genome (Fig. 2b). Each assembly had an average of 36 genes with a gain in copy number relative to GRCh38 within its predicted reliable regions, with a bias towards rare, low-copy CNVs (Fig. 2c). In detail, 71% of CNV genes appeared in a single haplotype. Previous studies using read depth found that rare CNVs generally occur outside regions annotated as being enriched in SDs[31]. The genome assemblies confirmed this observation in sequence-resolved CNVs. When stratifying duplicated genes on the basis of AF into singleton (present in one haplotype), low frequency (<10%) and high frequency, 15% (118 out of 771) of the singleton CNVs mapped to SDs as annotated in GRCh38. Duplicated genes with a higher population frequency had a greater fraction in SDs: 59% (83 out of 140) of low frequency and 81% (44 out of 54) of high frequency. Overall, 58 genes were CNVs in 10% or more of haploid assemblies, and 16 genes were amplified in the majority of individuals relative to GRCh38 (Fig. 2d and Supplementary Table 9). Many of these genes were individually highly copy-number polymorphic and part of complex tandem duplications (Fig. 2e). For example, *GPRIN2* is a copy-number polymorphic[32] based on read depth and has a sequence resolution of one to three additional copies duplicated in tandem in the pangenome (Fig. 2f). *SPDYE2* is similarly resolved as one to four additional copies duplicated in tandem (Fig. 2g). Other CNV genes were not contiguously resolved and reflect limitations of the current assemblies (see the associated article[33]). For example, the defensin gene *DEFB107A* has three to seven additional copies assembled across all samples; however, this gene was assembled into three to seven separate contigs that do not reflect the global organization of this gene.

## Constructing a draft pangenome

We used a sequence graph representation for pangenomes[12,14] in which nodes correspond to segments of DNA. Each node has two possible orientations, forward and reverse, and there are four possible edges between any pair of nodes to reflect all combinations of orientations (bidirected graph). The underlying haplotype sequences can be represented as walks in the graph. The model represents a generalized multiple alignment of the genome assemblies from which we built it, whereby haplotypes are aligned where they co-occur on a given node (Fig. 3a).

**Fig. 3: Pangenome graphs represent diverse variation.**




**a**, A pangenome variation graph comprising two elements: a sequence graph, the nodes of which represent oriented DNA strings and bidirected edges represent the connectivity relationships; and embedded haplotype paths (coloured lines) that represent the individual assemblies. **b**, Small variant sites in pangenome graphs stratified by the variant type and by the number of alleles at each site. MNP, multinucleotide polymorphism. **c**, SV sites in the pangenome graphs stratified by repeat class and by the number of alleles at each site. Other TE, a site involving mixed classes of transposable

elements (TEs). VNTR, variable-number tandem repeat, a tandem repeat with the unit motif length ≥7 bp. STR, short tandem repeat, a tandem repeat with the unit motif length ≤6 bp. Other LCR, low-complexity regions with mixed VNTR and STR and low-complexity regions without a clear VNTR or STR pattern. Other repeat, a site involving mixed classes of repeats. SegDup, segmental duplication. Low repeat, a small fraction of the longest allele in a site involving repeats. **d**, Pangenome minor AF (MAF) spectrum for biallelic SNP, VNTR, L1 and Alu variants in the MC and PGGB graphs. **e**,**f**, Number of autosomal small variants per sample (**e**) and SVs per haplotype (**f**) in the pangenome. Variants were restricted to the Dipcall-confident regions. Samples are organized by 1KG populations. **g**, Pangenome growth curves for MC (left) and PGGB (right). Depth measures how often a segment is contained in any haplotype sequence, whereby core is present in ≥95% of haplotypes, common is ≥5%. **h**, Small variants in the GIAB (v.3.0) 'easy' regions annotated with AFs from gnomAD (v.3.1.2).

The process of generating a combined pangenome representation is an active research area. The problem is nontrivial both because of computational challenges (there are hundreds of billions of bases of sequence to align) and because determining which alignments to include is not always obvious, particularly for recently duplicated and repetitive sequences. We applied three different graph construction methods that have been under active development for this project: Minigraph[34], Minigraph-Cactus (MC)[35] and PanGenome Graph Builder (PGGB)[36] (Extended Data Fig. 3 and Methods). The availability of these three models provided us with multiple views into homology relationships in the pangenome while supporting validation of discovered variation by independent methods. We included GRCh38 and T2T-CHM13 references within the pangenomes, and three samples (HG002, HG005 and NA19240) were held out to permit their use in benchmarking (hence 90 haplotypes total). In brief, Minigraph builds a pangenome by starting from a reference assembly, here GRCh38, and iteratively and progressively adds in additional assemblies, recording only SVs ≥ 50 bases. It admits complex variants, including duplications and inversions. MC extends the Minigraph pangenome with a base level alignment of the homology relationships between the assemblies using the Cactus genome aligner[37] while retaining the structure of the Minigraph pangenome. PGGB constructs a pangenome from an all-to-all alignment of the assemblies.

Although both T2T-CHM13 and GRCh38 are used to partition contigs into chromosomes, the PGGB graph is otherwise reference free (that is, it does not base itself on a chosen reference assembly).

## Measuring pangenome variation

The different algorithmic approaches used to construct a pangenome graph influence graph properties while representing the same underlying sequences. The basic properties of the three graphs produced with the different pangenome methods are shown in Supplementary Table 10. The Minigraph chart, by virtue of being limited to structural variation, is the smallest, with more than two orders of magnitude fewer nodes and edges than the base level graphs. Its length (3.24 Gb), measured as the total bases of all nodes, is similar to the MC graph (3.29 Gb) despite the latter adding many small variants. This difference is due to the MC graph also aligning a significant number of sequences left unaligned by Minigraph. The PGGB graph contains roughly 5 Gb more sequence because it includes highly structurally divergent satellite regions omitted from the other approaches (Methods) and does not implement any trimming or filtering of the input assembly contigs.

To characterize variants in the pangenome graphs, we used graph decomposition to identify ‘bubble’ subgraphs that correspond to non-overlapping variant sites. We then classified variant sites into small variants (<50 bp) and SVs (≥50 bp) of different types (Methods). We found similar numbers of each variant type in each pangenome, with 22 million small variants in the MC graph (21 million in PGGB) (Fig. 3b), and 67,000 SVs in the MC graph (73,000 in PGGB, 75,000 in Minigraph) (Fig. 3c). We assessed variation in each individual assembly by tracing their paths through the graphs and found similar numbers of small variants and SVs within confident genomic regions defined by Dipcall[38]. Specifically, there were 5.34 million small variants per sample and 16,800 SVs per haplotype on average in the MC graph (5.35 million and 17,400, respectively in PGGB) (Fig. 3e,f). Differences in variant counts among samples from different ancestry groups recapitulated previous observations[1]. There was a total of 90 Mb of non-reference sequence in the SV sites, excluding difficult-to-align centromeric repeats, in the MC graph (55 Mb for PGGB, 86 Mb for

Minigraph). Alu, L1 and ERV SVs appeared largely biallelic, whereas VNTRs frequently had three or more distinct alleles per site. The minor AF in the pangenomes of biallelic variants was similar for SNPs and for L1, Alu and VNTR variants, although VNTRs showed a slight shift towards more common alleles (Fig. 3d).

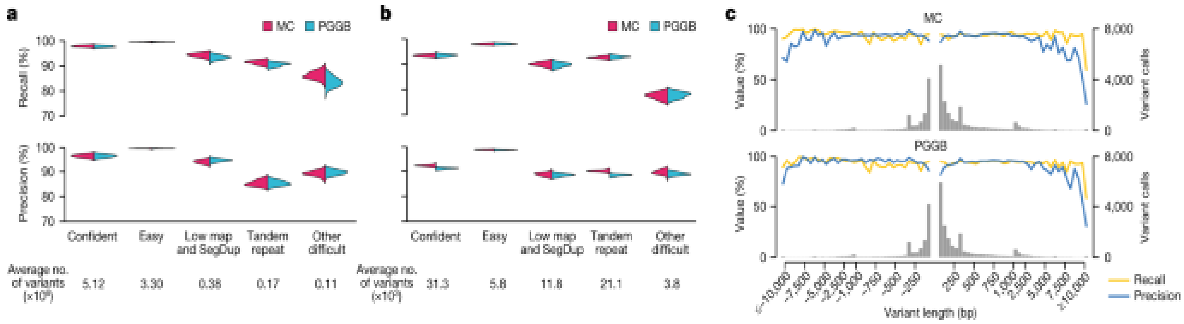
We quantified the amount of euchromatic autosomal non-reference (GRCh38) sequence that each of the 44 diploid genomes incrementally contributes to the pangenome (Fig. 3g and Methods) for both MC and PGGB graphs. We limited the analysis to the euchromatic sequence because we were generally confident in its assembly and alignment, and much of the heterochromatic sequence was omitted from the MC graph (Methods). Overall, the euchromatic autosomal non-reference sequence added up to about 175 Mb in the MC graph (around 190 Mb in PGGB), out of which about 55 Mb (around 105 Mb in PGGB) was observed only on a single haplotype. Our analysis further suggested that about 5 Mb and 70 Mb in the MC graph (around 10 Mb and 60 Mb in PCGB) could be attributed to core (present in ≥95% of all haplotypes) and common genomes (present in ≥5% of all haplotypes), respectively (Supplementary Table 11). We also estimated the growth of the euchromatic autosomal pangenome independent of the order of genomes by sampling 200 permutations (Supplementary Fig. 4) and recording the median pangenome size across all samples in the MC graph. Our results indicated that the second genome added around 23 Mb of euchromatic autosomal sequence to the pangenome, whereas the last genome tended to add only about 0.64 Mb. These numbers are conservative owing to additional highly polymorphic sequence residing in assembly gaps. Extrapolating under Heaps' Law[39] (Methods), we anticipate that at least an additional 150 Mb of euchromatic autosomal sequence will be added in the pangenome graph when HPRC produces 700 haplotypes in the future.

We annotated the small variants overlapping the GIAB (v.3.0) 'easy' regions (covering 74.35% of GRCh38) with AFs from gnomAD (v.3.1.2) (Fig. 3h and Supplementary Table 12). In the MC graph, about 60.2% (around 9.7 million variants) had an AF of 1% or greater. About 35.7% were rare, having an AF less than 1% but above zero. About 1.7% were singleton. The remaining 2.4% were missing from gnomAD. Similar results were obtained using the PGGB graph by repeating this exercise with small variant calls detected by pairwise alignment of assemblies to GRCh38 using Dipcall[38]

and by calling small variants from the HiFi sequencing data using DeepVariant[40]. Given that 1KG samples are included in gnomAD, these missing variants are expected to be a mixture of false negatives in gnomAD and false positives in the pangenome.

To further explore the quality of variant calls captured by assembly and graph construction, we compared pangenome-decoded variants against GRCh38 to variant sets identified by conventional reference-based genotyping methods (Supplementary Fig. 5 and Methods). These reference-based call sets were generated from the PacBio HiFi reads and haplotype-resolved assemblies using the following different discovery methods: DeepVariant[40], PBSV[41], Sniffles[42] with Iris[43], SVIM[44], SVIM-asm[45], PAV[5] and the Hall-lab pipeline (Methods). For benchmarking small variants, we excluded regions that contained SVs detected or implied by the alignment of the haploid assemblies of that sample to GRCh38, as current benchmarking tools do not account for different representations of small variants inside or near SVs (Methods). Comparing small variants (Fig. 4a) and SVs (Fig. 4b) from the pangenomes to the reference-based sets, we observed a high level of concordance that varied, as expected, by the relative repeat content of the surrounding genome. Overall, variant calling performance was high in both the MC and PGGB graphs. For example, in relatively unique easy genomic regions constituting 75.42% of the autosomal genome, samples showed a mean of 99.64% recall and 99.64% precision for small variants in the MC graph. Meanwhile, in high-confidence regions (around 90% of autosomal genome), samples showed 97.91% recall and 96.66% precision (Fig. 4a). Performance was lower for SVs than for small variants (Fig. 4b), as expected, but was still strong. Variant calling performance was lower in highly repetitive genome regions (3.87% of autosomal genome; Fig. 4a,b), for which more work will be required to achieve high-quality variant maps. These values are likely to be significant underestimates of variant calling quality, considering known errors in the truth set owing to the inherent limitations of reference-based variant callers (see below). Stratifying the insertion and deletion SVs within the pangenome, we observed relatively high levels of agreement with the reference-based methods regardless of length (Fig. 4c).

**Fig. 4: Pangenome graph evaluation.**




**a,b**, Precision and recall of autosomal small variants (**a**) and SVs (**b**) in the pangenomes relative to consensus variant sets. Small variants are compared to HiFi–DeepVariant calls. SVs are compared to the consensus of six reference-based SV callers (Methods). Comparisons are restricted to the Dipcall-confident regions and then stratified by the GIAB (v.3.0) genomic context. **c**, Average SV precision, recall and frequency in the Dipcall-confident regions stratified by length in the MC (top) and PGGB (bottom) graphs relative to consensus SV sets. The histogram bin size is 50 bp for SVs <1 kb and 500 bp for SVs ≥1 kb.

An independent measure of the quality of the pangenome graphs is the extent to which sample haplotype paths through the graph are well supported by the raw sequencing data. When we calculated the number of supporting reads by aligning them to the MC graph using GraphAligner (Methods), more than 97% of HiFi reads were aligned to the MC graph after filtering (Extended Data Fig. 4, left). We further calculated the read depth of on-target and off-target edges based on the sample paths in the graph. On average, more than 94% of on-target edges were supported by at least 5 reads, and we observed 2 peaks in the read depth distribution of on-target edges (Extended Data Fig. 4, middle): a minor peak corresponding to the edges in heterozygous regions, and a major peak at twice the minor peak corresponding to the edges in homozygous regions. By contrast, only 7% or fewer off-target edges were supported by at least 5 reads (Extended Data Fig. 4, right). In addition to HiFi reads, we used ONT reads from 29 out of the 44 samples to perform the same analysis. Even though the data were lower in coverage, similar results were obtained (Supplementary Figs. 6 and 7).

These data also show that the pangenome graphs performed better at capturing genome variation than the above benchmarking results imply. For

example, a mean of 89.3% of putative false-positive small variant calls were supported by ≥5 HiFi reads, and 75.3% by ≥10 reads (85.9% and 73.8%, respectively, for SVs). This result suggests that most putative errors are in fact real variants that were missed by the reference-based callers used to create the truth set (Supplementary Fig. 8 and Supplementary Table 13).

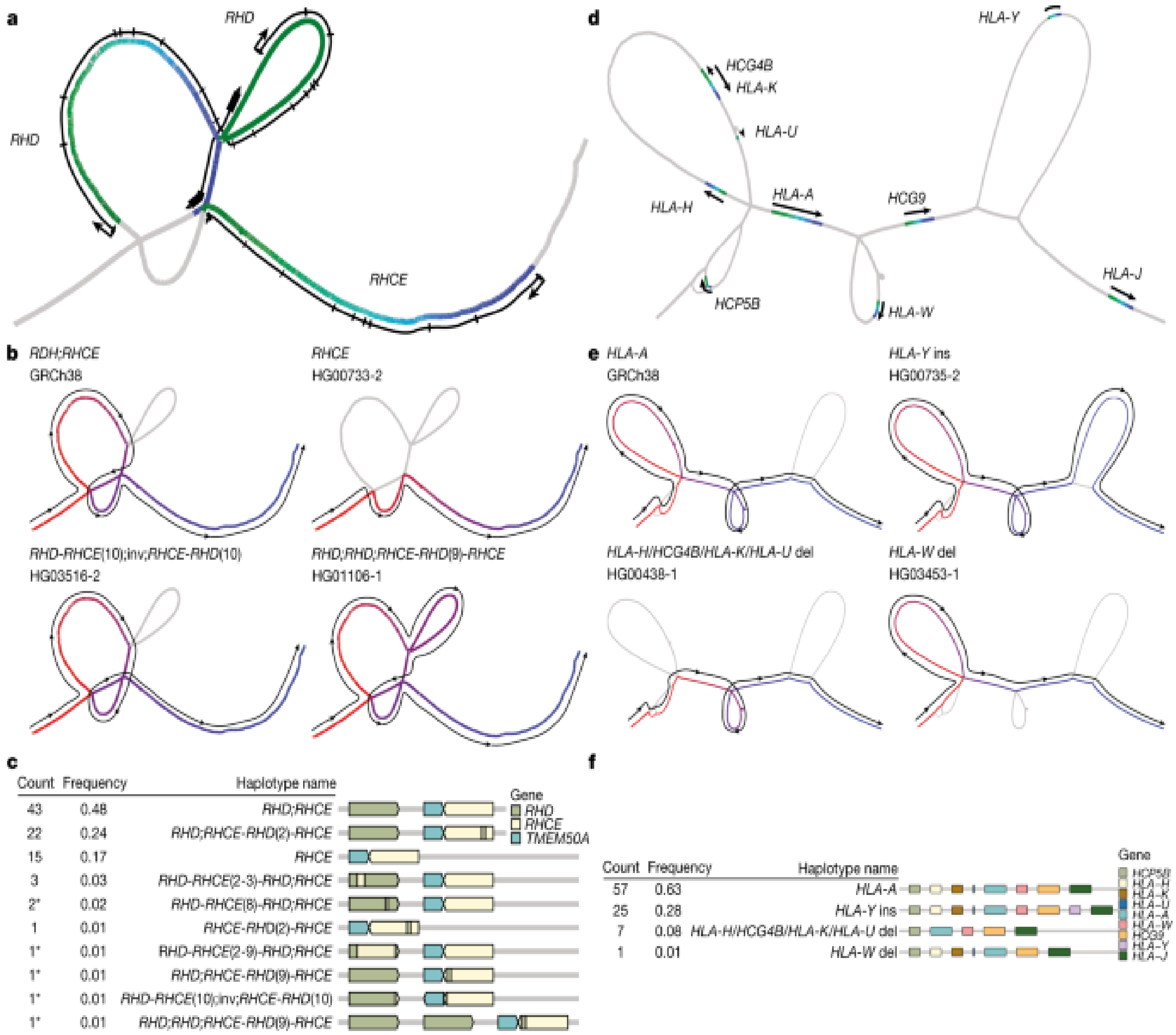
To assess gene alignments in the pangenome, we used the Comparative Annotation Toolkit (CAT)[46] to liftover GENCODE (v.38) annotations using the MC pangenome alignment onto the individual haplotype assemblies. CAT lifted and annotated a median of 99.1% of 86,757 protein-coding transcripts per assembly (Extended Data Fig. 5, Supplementary Fig. 9, Supplementary Tables 14 and 15 and Methods), making it comparable to the Ensembl-mapping-based pipeline (median of 99.4% per assembly). This result supports the idea that the MC pangenome captures most transcript homologies. When comparing the CAT and Ensembl annotations per assembly, median Jaccard similarities of 0.99 for both genes and transcripts were obtained (Methods). A median of 360 (0.4%) protein-coding transcripts per assembly mapped at different loci between the Ensembl and CAT annotations.

## Pangenomes represent complex loci

We next turned our attention to complex multiallelic SVs, which have historically been difficult to map using reference-based methods. To screen for complex SVs, we identified bubbles >10 kb from Minigraph that exhibited at least five structural alleles among the assembled haplotypes (Methods). We found that 620 out of 76,506 total sites (0.81%) were complex, and 44 of these overlapped with medically relevant protein-coding genes[47] (Supplementary Table 16). Some are well-known complex SV loci, and all are known to be structurally variable based on previous short-read SV mapping studies[10,19,32]. However, whereas previous short-read SV calls at these loci are typically imprecise owing to alignment issues and low-resolution read-depth analysis methods, here we resolved their structure at single-base resolution. We selected five clinically relevant complex SV loci for detailed structural analysis: *RHD–RHCE*, *HLA-A*, *CYP2D6–CYP2D7*, *C4* and *LPA* (Methods). For each locus and graph, we identified their locations within the graph and then annotated paths within this subgraph

with known genes. We traced the individual haplotypes through the subgraph to reveal the structure of each assembly. In *CYP2D6–CYP2D7* (Extended Data Fig. 6), *C4* (Supplementary Fig. 10) and *LPA* (Supplementary Fig. 11), we recapitulated previously described haplotypes. For *CYP2D6–CYP2D7*, our calls matched 96% of haplotypes of 76 assemblies called by Cyrius using Illumina short-read data[48]. Two discrepancies appeared to be caused by errors from Cyrius, and the third was a false duplication in the HG01071-2 pangenome assembly revealed by Flagger. This comparison suggests that the pangenomes faithfully agree with existing knowledge of this complex locus. In *RHD–RHCE* (Fig. 5a–c), in addition to previously described haplotypes, we inferred the presence of five new haplotypes, which included one duplication allele of *RHD* and one inversion allele between *RHD* and *RHCE* that swaps the last exon of both genes. Around *HLA-A* (Fig. 5d–f and Supplementary Fig. 12), two deletion alleles have been previously described—albeit with imprecise breakpoints[10]—but an insertion allele carrying a *HLA-Y* pseudogene was previously unreported. The long sequence (65 kb) inserted with *HLA-Y* occurred at high frequency (28%) but has little homology to GRCh38.

**Fig. 5: Visualizing complex pangenome loci.**




**a**–**c**, Structural haplotypes of *RHD* and *RHCE* from the MC graph. Locations of *RHD* and *RHCE* within the graph (**a**). The colour gradient is based on the precise relative position of each gene; green, head of a gene; blue, end of a gene. The lines alongside the graph are based on the approximate position of gene bodies, including exons and transcription start sites. Different structural haplotypes take different paths through the graph (**b**). The colour gradient and lines show the path of each allele; red, start of a path; blue, end of a path. Frequency and linear structural visualization of all structural haplotypes called by the graph among 90 haploid assemblies (**c**). Asterisks indicate newly discovered haplotypes. **d**–**f**, Structural haplotypes of *HLA-A* from the PGGB graph, visualized using the same conventions as **a**–**c**. del, deletion; ins, insertion; inv, inversion.

We also compared the representation of these five loci in the MC and PGGB graphs (Supplementary Fig. 13). Each graph independently recapitulated the same haplotype structures. In general, in the PGGB graph, many SV hotspots, including the centromeres, were transitively collapsed into loops through a subgraph representing a single repeat copy. This feature tends to reduce the size of variants found in repetitive sequences. Assemblies that contained multiple copies of the homologous sequence traversed these nodes a corresponding number of times. By contrast, the MC graph maintained separate copies of these homologous sequences.

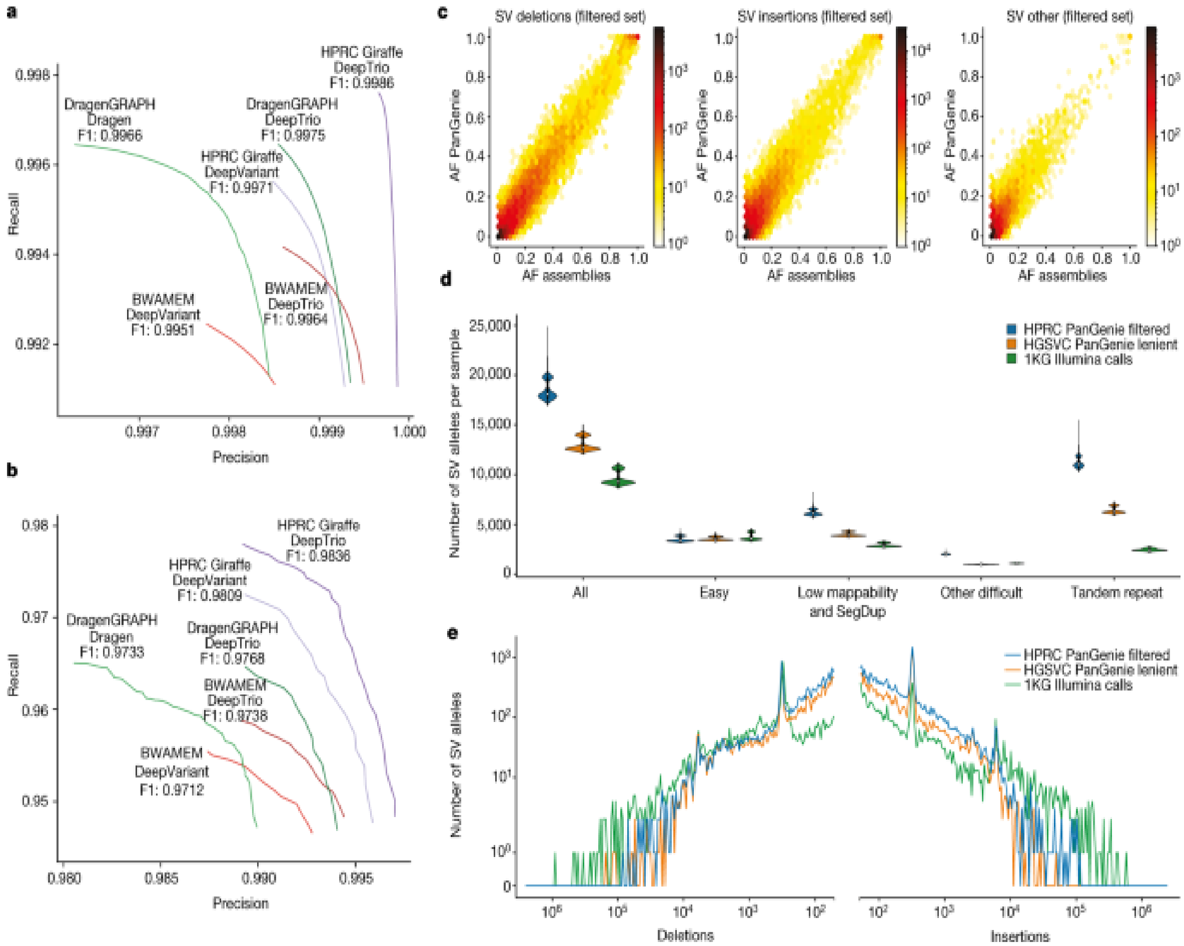
# Applications of the pangenome

## Pangenome-based short variant discovery

Our pangenome reference aims to broadly improve downstream analysis workflows by removing mapping biases that are inherent in the use of a single linear reference genome such as GRCh38 or CHM13. As an initial test case, we studied whether mapping against our pangenomes could improve the accuracy of calling small variants from short reads. We used Giraffe[49] to align short reads from the GIAB benchmark samples[18] to the MC pangenome graph. For comparison, we aligned reads to GRCh38 using BWA-MEM[50] and to Dragen Graph[51], which uses GRCh38 augmented with alternative haplotypes at variant sites. We called SNPs and indels with DeepVariant[40] and the Dragen variant caller[51] (Methods). Our pangenomic approach (Giraffe plus DeepVariant) outperformed the other approaches for calling small variants (Fig. 6a), with gains for both SNPs and indels (Supplementary Fig. 14 and Supplementary Table 17). For example, it made 21,700 errors (false positives or false negatives) in the confident regions of the GIAB truth set using 30× reads from HG005. By contrast, 36,144 errors were made when DeepVariant used the reads aligned to GRCh38, and 26,852 errors when using the Dragen pipeline. In challenging medically relevant genes[47], the increase in performance was even larger for both SNPs (F1 score, defined as the harmonic mean of precision and recall, of 0.985 for Giraffe plus DeepVariant compared with <0.976 for the other methods) and indels (F1 score of 0.961 for Giraffe plus DeepVariant compared with <0.958 for the other methods) (Fig. 6b). Many regions benefitted from using

pangenome mapping, but regions with errors in GRCh38 and large L1HS sequences benefitted the most from the pangenomic approach (Supplementary Figs. 15 and 16).

**Fig. 6: Performance gains for pangenome-aided analysis of short-read WGS data.**




**a**,**b**, Precision–recall curves showing the performance of different combinations of linear reference and various mappers and variant callers evaluated against the GIAB (v.4.2.1) HG005 benchmark (**a**) and the challenging medically relevant genes (CMRG; v.1.0) benchmark (**b**). Giraffe uses the MC pangenome graph, BWA-MEM uses GRCh38 and Dragen Graph uses GRCh38 with additional alternative haplotype sequences. **c**, Comparison of AFs observed from the PanGenie genotypes for all 2,504 unrelated 1KG samples and the AFs observed across 44 of the HPRC assembly samples in the MC graph. The PanGenie genotypes include all variants contained in the filtered set (28,433 deletions, 84,755 insertions,

32,431 other alleles). **d**, Number of SVs present (genotype 0/1 or 1/1) in each of the 3,202 1KG samples in the filtered HPRC genotypes (PanGenie) after merging similar alleles ($n$ = 100,442 SVs), the HGSVC lenient set ($n$ = 52,659 SVs) and the 1KG Illumina calls ($n$ = 172,968 SVs) in GIAB regions. In the box plots, lower and upper limits represent the first and third quartiles of the data, the white dots represent the median and the black lines mark minima and maxima of the data points. **e**, Length distribution of SV insertions and SV deletions contained in the filtered HPRC genotypes (PanGenie), the HGSVC lenient set and the 1KG Illumina calls. Only variants with a common AF > 5% across the 3,202 samples were considered.

We next benchmarked variant calling using parent–child trios. Using DeepTrio[52] resulted in better performance compared with DeepVariant across all samples of the GIAB (Fig. 6a) and the challenging medically relevant gene benchmarks (Fig. 6b and Supplementary Fig. 14). Moreover, improvements appeared to be additive to those from the pangenome. For example, DeepTrio using Giraffe alignments gave the highest calling accuracy, with the number of errors decreasing from 21,700 (single sample calling) to 10,098 (trio calling) for HG005.

## A pangenome variant resource

To create a community resource to aid the development of methods and the analyses of pangenome-based population genetics, we used Giraffe to align high-coverage short-read data from 3,202 samples of the 1KG[19] to our pangenome graph and DeepVariant to call small variants (Methods). The Mendelian consistency computed across 100 trios from those samples was comparable to the one computed across samples from the GIAB truth set, which indicated that comparable call set quality was obtained (Supplementary Figs. 17 and 18). The number of small variants called was consistently higher across different ancestries, with on average 64,000 more variants per sample compared with the 1KG catalogue (Extended Data Fig. 7a). Given that our pangenome-based calls showed improved performance in challenging regions (Fig. 6b), this call set across the 1KG cohort now provides the genetics and genomics communities with AF estimates for complex but medically relevant loci. For example, our approach was able to detect the gene conversion event covering the second exon *RHCE*, which

was observed in about 25% of assembled haplotypes (Fig. 5c and Extended Data Fig. 8). Moreover, for *KCNE1*, we provide calls and frequencies in a 40 kb region, spanning 3 exons, that could not be previously assessed owing to the presence of a false duplication in GRCh38 (Supplementary Fig. 19; see also an associated article[53] for genome-wide analysis of interlocus gene conversion).

## SV genotyping

The ability to represent polymorphic SVs is a key advantage of a graph-based pangenome reference. To demonstrate the utility of the sequence-resolved SVs inherent to our pangenome, we used PanGenie[54] to genotype the bubbles in the MC graph. We decomposed bubbles into their constituent variant alleles (Supplementary Figs. 20 and 21) and found that 22,133,782 bubbles represented 20,194,117 SNP alleles, 6,848,115 indel alleles and 413,809 SV alleles (Supplementary Fig. 22 and Methods). Of these SV alleles that were non-reference (neither GRCh38 nor T2T-CHM13), 17,720 were observed in biallelic contexts and 396,089 at multiallelic loci with more than 1 non-reference allele, including extreme cases in which all 88 haplotypes showed distinct alleles (Supplementary Fig. 22). To analyse the genotyping performance of PanGenie, we conducted a leave-one-out experiment in which we repeatedly removed one sample from the graph and re-genotyped it using the remaining haplotype paths in the graph and short-read data for the left-out sample (Methods). In line with previous results[5,54], we obtained high genotype concordance across all variant types and genomic contexts (Extended Data Fig. 9). Furthermore, we used PanGenie to genotype HG002 and evaluated genotypes based on SVs at challenging medically relevant loci[47]. This analysis resulted in a precision of 0.74 and an adjusted recall of 0.81 (Methods).

Next we genotyped the 3,202 samples from the 1KG[19] (Methods). We filtered the resulting SV genotypes using a machine-learning approach[5,54] that assessed different statistics, including Mendelian consistency and concordance, to assembly based calls. As a result, we produced a filtered, high-quality subset of SV genotypes containing 28,434 deletion alleles, 84,752 insertion alleles and 26,439 other SV alleles (Supplementary Table 18 and Methods). Many of the alleles not included in the filtered set

stemmed from complex, multiallelic loci and were enriched for rare alleles. As independent quality control measures for genotypes in the filtered set, we assessed the Hardy–Weinberg equilibrium values (Supplementary Figs. 23–25) and compared AFs observed across the genotypes of all 2,504 unrelated samples to the respective AFs of the 44 assembly samples (88 haplotypes) contained in the graph. Pearson correlation values of 0.96, 0.93 and 0.90 for the deletion, insertion and other SV alleles, respectively, were observed (Fig. 6c), which indicated the high quality of the genotypes. To quantify our ability to detect additional SVs, we compared our filtered set of genotypes to the HGSVC PanGenie genotypes (v.2.0 'lenient' set)[5] and the Illumina-based 1KG SV call set[19]. We analysed the number of detected SV alleles in each sample (homozygous or heterozygous) and stratified them by genome annotations from GIAB (Fig. 6d and Methods) as well as using our own more detailed annotations (Supplementary Fig. 26). Both of the PanGenie-based call sets detected more SVs (HPRC, 18,483 SVs per sample; HGSVC, 12,997 SVs per sample) than the short-read-based 1KG call set (9,596 SVs per sample), with a particularly substantial advance for deletions <300 bp and insertions (Fig. 6e). The respective average numbers of SVs per haplotype were 12,439 for HPRC, 9,227 for HGSVC and 6,099 for the 1KG calls (Supplementary Fig. 27); that is, a gain of 104.0% for HPRC over 1KG and of 34.8% over HGSVC. This result confirms that short-read-based SV discovery relative to a linear reference genome misses a large proportion of SVs[5,6,8]. As anticipated, the number of SVs per sample within 'easy' genomic regions was consistent across all three call sets, particularly in low-mappability and tandem repeat regions, and the use of our pangenome reference led to substantial gains (Fig. 6d), including for common variants (Fig. 6e and Supplementary Fig. 28). Although the newly identified SVs were harder to genotype because they are primarily located in repetitive regions, genotype concordances were high and close to the ones for known SVs (Supplementary Fig. 29).

## Improved tandem repeat representation

VNTRs are particularly variable among individuals and are challenging to access with short reads. The gains in the number of genotyped SVs in VNTRs (Fig. 6d and Supplementary Fig. 28) prompted us to investigate whether our pangenome reference could also improve read mapping in

VNTR regions. We first established orthology mapping between haplotypes in our pangenome reference using danbing-tk[55]. The orthology can be established for 94,452 out of the 98,021 VNTR loci (96.4%) discovered by TRF[56]. When mapping simulated short reads to GRCh38 with BWA-MEM, the rate of unmapped reads was 6.6–8.5 times greater compared with mapping to the MC graph with Giraffe (Extended Data Fig. 10a, Supplementary Fig. 30 and Supplementary Table 19). The true negatives were on average 1.9% higher than the GRCh38 approach, and the true positives were on average 0.087% higher. The graph approach also reduced false negatives by 2.1-fold. Read depth over a locus is correlated with the copy number of a duplication and we evaluated how well length variants in VNTR regions can be estimated using either the MC graph or GRCh38. The graph approach performed better for 80% of the loci (48,085 out of 60,386) and increased the median $r^2$ from 0.58 to 0.70 (Supplementary Fig. 31).

## Improved RNA sequencing mapping

To evaluate the benefit of our pangenome reference on transcriptomics, we simulated RNA sequencing (RNA-seq) reads and mapped them to a pangenome and to a standard reference genome (Methods). The pangenome-based pipeline using vg mpmap[57] achieved significantly lower false mapping rates than a linear reference pipeline using either vg mpmap or STAR[58] (Extended Data Fig. 10b). Compared with the linear reference pipelines, the pangenome pipeline also showed reduced allelic bias and increased mapped coverage on heterozygous variants, which could benefit studies of allele-specific expression (Supplementary Fig. 32). With real sequencing data, mapping rates were more difficult to interpret in the absence of a ground truth (Supplementary Fig. 33). Instead, we focused on the correlation in exon coverage to independent PacBio long-read Iso-Seq data. The analysis showed that the correlation was highest when mapping to a spliced pangenome graph derived from the MC graph (Supplementary Fig. 34). The pangenome pipeline showed a modest increase in correlation over the linear reference pipelines (0.006–0.011). In addition, mapping the simulated reads to the MC graph led to improved gene expression estimates relative to the linear GRCh38, regardless of whether alternative contigs were included in GRCh38 (Supplementary Fig. 35).

### Improved chromatin immunoprecipitation and sequencing analysis

We used the pangenome to re-analyse H3K4me1 and H3K27ac data from chromatin immunoprecipitation and sequencing (ChIP-seq) and assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) of monocyte-derived macrophages from 30 individuals with African ancestry or European ancestry[59]. Overall, we observed a net increase in the number of peak calls, whereby, on average, 2–3% of peaks were found only when using the MC pangenome (Extended Data Fig. 10c). Moreover, the newly found peaks were replicated in more samples than expected by chance (Supplementary Fig. 36). In addition, we recovered epigenomic features that were specific to SV alleles not present in GRCh38 (termed non-reference). For example, across all H3K4me1 datasets, we assigned 1,326 events to the non-reference SV allele, 1,443 to the reference allele and 2,008 to both alleles within heterozygous SVs (Extended Data Fig. 10d), with some replicated multiple times across samples (Supplementary Fig. 37). Of these, there were 194 SVs with peaks that were observed only in African ancestry genomes, 150 that were observed only in European ancestry genomes and 216 that were observed in both African and European ancestry genomes. As expected, rare alleles were enriched for ancestry-specific events (Supplementary Fig. 37).

## Discussion

We have publicly released 94 de novo haplotype assemblies from a diverse group of 47 individuals. This provides a large set of fully phased human genome assemblies and outperforms earlier efforts on many levels of assembly quality[5,16,60]. For example, compared with Ebert et al.[5], the average median base level accuracy is nearly an order of magnitude higher, the N50 contiguity is nearly double and the structural accuracy is higher[33]. These improvements are the result of recent improvements in de novo assembly driven both by better sequencing technology and coordinated innovations in assembly algorithms[20,21]. To validate assembly structural accuracy, we developed a new pipeline that maps low error, long reads to each diploid assembly to support the predicted haplotypes. This pipeline

indicated that more than 99% of each assembly, and greater than 90% of the assembled sequence representing highly repetitive arrays, was structurally correct. Some challenges around loci that harbour copy number polymorphisms and/or inversions remained[33]. Although the focus of this effort was to build a reference resource, highly accurate haplotype-resolved assemblies enabled us to access previously inaccessible regions, highlighting new forms of genetic variation and providing new insights into mutational processes such as interlocus gene conversion[53].

Accompanying these assemblies are 94 sets of Ensembl gene annotations, representing a large collection of de novo assembled human transcriptome annotations. Each transcriptome annotation is highly complete, particularly for protein-coding transcripts. These putative transcriptome annotations enabled us to analyse sequence-resolved CNVs. In detail, we assembled genic CNVs (mostly singletons) for 1,115 different protein-coding genes, confirming earlier mapping-based analyses that predicted that the majority of rare genic CNVs occur outside known SDs[31]. These CNV genes accounted for 0.6–4.4 Mb of additional genic sequences per haplotype compared with GRCh38. These contained genes known to have CNVs associated with human health, including amylase[61] (four to ten copies), β-defensin[62] (three to seven copies, *DEFB107A*) and *NOTCH2NLC*–*NOTCH2NLB*[63] (one additional copy).

The pangenomes presented here are both a set of individual haploid genome assemblies and an alignment of these assemblies. The combination can be efficiently described as a variation graph[14,64]. A new set of exchange formats for pangenomics, including extensions of the graphical fragment format (GFA) that encode variation graphs, are emerging[34]. An associated article[65] to this work demonstrated that the pangenomes presented here can be losslessly stored using a compressed, binary representation of GFA in just 3–6 GB despite representing more than 282 billion bases of individual sequence, with strongly sublinear scaling as new genomes are added. Creating pangenome graphs is an active research topic, so we developed multiple pipelines, and details of these methods are further explored in companion papers[35,36]. We found concordance between the different construction approaches used here, whereby the MC and PGGB pangenomes contained nearly the same number of small variants and SVs of various

types. Furthermore, these encoded pangenome variants showed high levels of agreement with existing linear reference-based methods for variant discovery, particularly within the non-repetitive fraction of the genome. Our study of complex and medically relevant loci showed that the pangenomes faithfully recapitulated existing knowledge and will enable future efforts to study the role of complex variation in human disease. Further work will be required to more comprehensively identify medically relevant complex SVs and to ensure the accuracy of each allele represented in the pangenome.

Where the pangenome graphs differ is principally in how they handle CNV sequences. The PGGB method will frequently merge CNVs, whereas the MC graphs represent CNV copies as independent subgraphs. Both approaches have merits, and which approach to favour will take further experimentation and community input, and may vary by the specific application. The PGGB method retained all centromeric and satellite sequences, whereas the MC graph pruned much of this sequence. This made it practical with current methods to use the MC graphs for read alignment applications. However, pruning these sequences is not a satisfactory solution. Longer term, more work is needed to determine how best to align and represent these large repeat arrays within pangenomes, particularly as T2T assembly becomes commonplace and these arrays therefore completed. Furthermore, although the PGGB graph retained centromeric and satellite sequences, in principle, by enabling analysis of previously inaccessible parts of the pangenome, our initial population-genetic analysis of these regions (Methods) leaves open questions about assembly accuracy and alignment, especially in areas of the genome where mutation rates are thought to be an order of magnitude greater[66]. This suggests that significant care must be taken when studying them, and new methods may need to be developed to fully understand and characterize this component of the human pangenome.

A near-term application of pangenome references will be to improve reference-based sequence mapping workflows. In these workflows, the pangenome can act as a drop-in replacement for existing references, with the read mappings projected from the pangenome space back onto an existing linear reference for downstream processing. This is how the Giraffe–DeepVariant workflow functions: DeepVariant, the variant caller, never needs to consider the complexity of the pangenome, but the workflow benefits from a mapping step that accounts for sequences that are missing

from the linear reference. Making the switch to using pangenome mapping is not significantly more computationally expensive[49] and resulted in an average 34% reduction in false-positive and false-negative errors compared with using the standard reference methods (Supplementary Fig. 38). These benefits were also greatest at complex loci[47]. Pangenomes not only improve variant calling but also improve transcript mapping accuracy[57] and detection of ChIP-seq peaks[67].

SVs have been mostly excluded from short-read studies because methods to genotype them using a linear reference have limited accuracy and sensitivity. Previous short-read, linear reference studies have discovered 7,500–9,500 SVs per sample[19,68], whereas long-read sequencing efforts have routinely discovered around 25,000. Ebert et al.[5] showed that using PanGenie, a pangenomic approach, with 32 samples, a subset of these variants could be genotyped in short-read genomes (about 13,000 genotyped on average, ranging from 12,000 to 15,000 per sample). Using the same PanGenie method, the HPRC pangenome increases this to around 18,500 (ranging from 16,900 to 24,900) per sample, enabling the genotyping of the substantial majority of SVs discovered using long-reads per sample. The draft pangenome therefore delivers better SV calling than previous approaches, extracting latent information from short-read samples that are already available. So, in the future, the pangenome will enable the inclusion of tens of thousands of additional SV alleles into genome-wide association studies. Looking beyond short reads, in the future, the combination of the pangenome and low-cost long-read sequencing should prove to be a potent combination for comprehensive SV genotyping.

These new pangenomic workflows could benefit individuals of different ancestries differently. For read mapping and small variant calling, we observed a consistent improvement across individuals (Extended Data Fig. 7). Moreover, the pangenome might improve SV genotyping differently across individuals owing to the stronger divergence of the alleles from the reference. In the 1KG cohort, we observed that the genotyped samples clustered by super-population labels (Extended Data Fig. 11), which would suggest different levels of detection bias that are mitigated with the pangenome. However, we caution that the composition of the samples underlying the pangenome relative to the composition of the set of samples

genotyped could potentially influence these results; an analysis with more samples is warranted.

The openly accessible, diverse assemblies and pangenome graphs we present here form a draft of a pangenome reference. There are many remaining challenges to growing and refining this reference. For example, assembly reliability analysis revealed roughly an order of magnitude more erroneously assembled sequences in the HPRC assemblies than in the T2T-CHM13 complete assembly. Similarly, in a companion analysis, Strand-seq data from a subset of assemblies revealed 6–7 Mb of incorrectly oriented sequence per haplotype[33], which indicates that there is room to structurally improve the assemblies. Furthermore, despite being predicted to have less than 1 base error per around 200,000 assembled bases, base level sequencing errors are still an issue. For example, we identified more than a dozen apparent frameshifts and nonsense mutations per genome annotation that are probably the result of sequencing errors. The cohort we present is also relatively small notwithstanding the significant effort to generate the underlying long-read sequencing resource. Our near-term goal is to expand the pangenome to a diverse cohort of 350 individuals (which should capture most common variants), to push towards T2T genomes for this cohort (to properly represent the entire genome in almost all individuals) and to refine the pangenome alignment methods (so that telomere-to-telomere alignment is possible, capturing more complex regions of the genome). This will give us a more comprehensive representation of all types of human variation.

We acknowledge that references generated from the 1KG samples alone are insufficient to capture the extent of sequence diversity in the human population. To ensure that we are able to maximize our surveys of sample diversity while abiding by principles of community engagement and avoiding extractive practices[14,15], we will broaden our efforts to recruit new participants to improve the representation of human genetic diversity. A richer human reference map promises to improve our understanding of genomics and our ability to predict, diagnose and treat disease. A more diverse human reference map should also help ensure that the eventual applications of genomic research and precision medicine are effective for all populations. We recognize that the value of this project will partly be in the future establishment of new standards for how we capture variant diversity, the opportunity to disseminate science into diverse communities and

continued efforts to engage with diverse voices in this ambitious goal to build a common global reference resource. The methods we are developing should prove valuable for other species. Indeed, other groups are pioneering such efforts[69,70]. In parallel with our efforts to obtain a more comprehensive collection of diverse and highly accurate human reference genomes, we anticipate further optimization and rapid improvement of the pangenome reference, enabling an increasingly broad set of applications and use cases for both the research and clinical communities.

# Methods

## Sample selection

We identified parent–child trios from the 1KG in which the child cell line banked within the NHGRI Sample Repository for Human Genetic Research at the Coriell Institute for Medical Research was listed as having zero expansions and two or fewer passages, and rank-ordered representative individuals as follows. Loci with MAFs less than 0.05 were removed. MAFs were measured in the full cohort (that is, 2,504 individuals, 26 subpopulations) regardless of each individual's subpopulation labelling. For each chromosome, principal component analysis (PCA) was performed for dimension reduction. This resulted in a matrix with 2,200 features, which was then centred and scaled using smartPCA normalization. The matrix was further reduced to 100 features through another round of PCA.

We defined the representative individuals of a subpopulation as those who are similar to the other members in the group (which, in this scenario, is the subpopulation they belong to), as well as different from individuals outside the group. Group is defined by previous 1KG population labels (for example, 'Gambian in Western Division'). We did this as follows. For each sample, we first calculated the intragroup distance $d_{intra}$, which is the average of L2-norms between the sample and samples of the same subpopulation. The intergroup distance, $d_{inter}$, was similarly defined as the average of L2-norms between the sample and samples from all other subpopulations. The L2-norms were derived in the feature space of the PCA. The score of this sample was then defined as $10 \times d_{intra} + d_{inter}/(n - 1)$, where

$n$ is the number of subpopulations. For each subpopulation, if fewer than three trios were available, all were selected. Otherwise, trios were sorted by ranking children with max($paternal_{rank}$, $maternal_{rank}$), where $paternal_{rank}$ and $maternal_{rank}$ are the respective ranks of each parent's score, selecting the three trios with a maximum value. We ranked by parent scores because during the year 1 effort, the child samples did not have sequencing data and therefore had to be represented by the parents.

Ideally, we would have selected the same number of candidates from each subpopulation and have an equal number of candidates from both sexes. To correct for imbalances, we applied the following criteria for each subpopulation's candidate set: (1) when the sex was unbalanced (that is, off by more than one sample), we tried to swap in the next-best candidate of the less represented sex or did nothing if this was not possible; (2) if a subpopulation had fewer individuals than the desired sample selection size (that is, all candidates were selected), their unused slots were distributed to other unsaturated subpopulations. The latter choice is arbitrary but should have little impact on the overall results.

The genetic information used in this study was derived from publicly available cell lines from the NHGRI Sample Repository for Human Genetic Research and the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. Therefore, this study is exempt from human research approval as the proposed work involved the collection or study of data or specimens that are already publicly available.

## Sequencing

### Cell line expansion and banking

Lymphoblastoid cell lines (LCLs) used for sequencing from the 1KG collection (Supplementary Table 1) were obtained from the NHGRI Sample Repository for Human Genetic Research at the Coriell Institute for Medical Research. HG002 (GM24385) and HG005 (GM24631) LCLs were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. All expansions for sequencing were derived from the original expansion culture lot to ensure the lowest possible number of

passages and to reduce overall culturing time. Cells used for HiFi, Nanopore, Omni-C, Strand-seq, 10x Genomics and Bionano production and for g-banded karyotyping and Illumina Omni2.5 microarray were expanded to a total culture size of $4 \times 10^8$ cells, which resulted in a total of five passages after cell line establishment. Cells were split into production-specific sized vials as follows: HiFi, $2 \times 10^7$ cells; Nanopore, $5 \times 10^7$ cells; Omni-C, $5 \times 10^6$ cells; Strand-seq, $1 \times 10^7$ cells; 10x Genomics, $4 \times 10^6$ cells; and Bionano, $4 \times 10^6$ cells. Cells for Strand-seq were stored in 65% RPMI-1640, 30% FBS and 5% DMSO and frozen as viable cultures. All other cells were washed in PBS and flash-frozen as dry cell pellets. Cells used for ONT-UL production were separately expanded from the original expansion culture lot to a bank of five vials of $5 \times 10^6$ cells. A single vial was subsequently expanded to a total culture size of $4 \times 10^8$ cells, which resulted in a total of eight passages. Cells were also reserved for g-banded karyotyping and Illumina Omni2.5 microarray.

## Karyotyping and microarray

G-banded karyotype analysis was performed on $5 \times 10^6$ cells collected at passage five (for HiFi, Nanopore and Omni-C) and passage eight (for ONT-UL). For all cell lines, 20 metaphase cells were counted, and a minimum of 5 metaphase cells were analysed and karyotyped. Chromosome analysis was performed at a resolution of 400 bands or greater. A pass/fail criterion was used before cell lines proceeded to sequencing. Cell lines with normal karyotypes (46,XX or 46,XY) or lines with benign polymorphisms that are frequently seen in apparently healthy individuals were classified as passes. Cell lines were classified as failures if two or more cells harboured the same chromosomal abnormality. DNA used for microarray was isolated from frozen cell pellets ($3 \times 10^6$ to $7 \times 10^6$ cells) using a Maxwell RSC Cultured Cells DNA kit on a Maxwell RSC 48 instrument (Promega). DNA was genotyped at the Children’s Hospital of Philadelphia’s Center for Applied Genomics using an Infinium Omni2.5-8 v.1.3 BeadChip (Illumina) on an iScan System instrument (Illumina).

## HiFi sequencing

PacBio HiFi sequencing was distributed between two centres: Washington University in St. Louis and the University of Washington. We describe the protocols used at each centre separately.

**Washington University in St. Louis**

High-molecular-weight DNA was isolated from frozen cell pellets using a Qiagen MagAttract HMW DNA kit and sheared using a Diagenode Megaruptor I to 20 kb mode size. At all steps, DNA quantity was checked on a Qubit Fluorometer I with a dsDNA HS Assay kit (Thermo Fisher), and sizes were examined on a FEMTO Pulse (Agilent Technologies) using a Genomic DNA 165 kb kit. SMRTbell libraries were prepared for sequencing according to the protocol 'Procedure & Checklist—Preparing HiFi SMRTbell Libraries using the SMRTbell Express Template Prep Kit 2.0'. After SMRTbell generation, material was size-selected on a SageELF system (Sage Science) using the '0.75% 1-18 kb' program (target 3,450 bp in well 12), and some combinations of fraction 3 (average size of 15–21 kb), fraction 2 (average size of 16–27 kb) and fraction 1 (average size of 20–31 kb) were selected for sequencing, depending on the empirical size measurements and available mass. The selected library fractions were bound with Sequencing Primer v.2 and Sequel II Polymerase v.2.0 and sequenced on Sequel II instruments (PacBio) on SMRT Cells 8M using Sequencing Plate v.2.0, diffusion loading, 2 h of pre-extension and 30 h of movie times. Samples were sequenced to a minimum HiFi data amount of 108.5 Gbp (35× estimated genome coverage) on four SMRT Cells.

**University of Washington**

High-molecular-weight DNA was isolated from frozen cell pellets using a modified Gentra Puregene method and sheared using gTUBE (Covaris) to 20 kb mode size. At all steps, DNA quantity was checked by fluorometry on a DS-11 FX instrument (DeNovix) with a Qubit dsDNA HS Assay kit (Thermo Fisher), and sizes were examined on a FEMTO Pulse (Agilent Technologies) using a Genomic DNA 165 kb kit. SMRTbell libraries were prepared for sequencing according to the protocol 'Procedure & Checklist—Preparing HiFi SMRTbell Libraries using the SMRTbell Express Template Prep Kit 2.0'. After SMRTbell generation, material was size-selected on a

SageELF system (Sage Science) using the ‘0.75% 1–18 kb’ program (target 3,400 bp in well 12), and fraction 2 (average size of 17–20 kb) or fraction 1 (average size of 18–20 kb) was selected for sequencing, depending on the empirical size measurements and available mass. For some samples, the SageELF program ‘0.75% agarose, 10 kb–40 kb’ (target 10,000 bp in well 10) was used, and fractions 6 and 7 were pooled together for sequencing (average size of 17–21 kb). The selected library fractions were bound with Sequencing Primer v.2 and Sequel II Polymerase v.2.0 and sequenced on Sequel II instruments (PacBio) on SMRT Cells 8M using Sequencing Plate v.2.0, diffusion loading, 3–4 h of pre-extension and 30 h of movie times. Samples were sequenced to a minimum HiFi data amount of 96 Gbp (30× estimated genome coverage) on at least four SMRT Cells.

### Comparisons of HiFi production methods

Although subtle differences in HiFi data production methods existed between the University of Washington and Washington University in St. Louis, the resulting data were highly similar, with overlapping assembly statistics from most samples. These initial genomes were sequenced at a time when methods were being refined and optimized for HiFi sequencing, as it was a relatively new process. The primary differences in protocols are part of the nucleic acid isolation, fragmentation and size selection, with the downstream sequencing-specific applications being more consistent. Both teams were closely engaged with each other as well as with our company associates, including New England Biolabs (NEB), Qiagen, Diagenode and Sage Science, to provide optimal end products.

### Nanopore ultra-long sequencing protocol

For the 18 additional samples, we used the nanopore unsheared long-read sequencing protocol[16]. This generated about 60× coverage of unsheared sequencing from 3 PromethION flow cells and a N50 value of around 44 kb. For the 29 newly selected HPRC samples (Results), we used the protocol outlined below.

### DNA extraction

Around 50 million cells in a pellet were resuspended in 200 µl of PBS, and the resuspended cells were aliquoted (40 µl) into five 1.5 ml DNA Lo-bind Eppendorf tubes. The following procedure for DNA extraction was completed for each of the five aliquots. Each tube contained sufficient DNA for three libraries loaded onto one flow cell. The following reagents were added in sequence to each tube with pipette mixing (10 times up and down) using a P200 wide-bore pipette: 40 µl of proteinase K, 40 µl of buffer CS and 40 µl of CLE3. The samples were then incubated at room temperature (18–25 °C) for 30 min. Next, 40 µl of RNase A was added to each tube with pipette mixing (10 times) with a P200 wide-bore pipette, and samples were incubated at room temperature for 3 min. Two hundred microlitres of BL3 was mixed with 200 µl PBS in a 1.5 ml Eppendorf tube. Four hundred microlitres of this BL3–PBS mixture was then added to each sample and the samples mixed 10 times with a P1000 wide-bore pipette set to 600 µl.

Samples were incubated for 10 min at room temperature and then pipette mixed 5 times, then incubated at room temperature for 10 min and pipette mixed 5 times and then further incubated for 10 min at room temperature. A white precipitate may form after addition of BL3. This is normal. A Nanobind disk was added to the cell lysate, then 600 µl of isopropanol was added. Mixing was performed by inversion of the tube 5 times. Tubes were further mixed on a tube rotator (9 r.p.m. at room temperature for 10 min). The tubes were then placed on a magnetic tube rack, and the Nanobind disk positioned closer to the top of the tube to avoid inadvertent removal of the DNA bound to the Nanobind disk. The supernatant was discarded using a pipette and 700 µl of buffer CW1 was added to each tube. The tube in the magnetic rack was then inverted 4 times for mixing. A second and third wash with 500 µl of buffer CW2 (inversion mix 4 times for each wash) was performed. After the second CW2 wash, liquid was removed from the tube cap and the tubes spun on a mini-centrifuge for 2 s, and replaced on the magnetic rack. Residual liquid was removed from the bottom of the tube, taking care not to remove DNA associated with the Nanobind disk. Elution from the Nanobind disk was accomplished by adding 160 µl Circulomics elution buffer (EB) plus 0.02% Triton X-100 (comprising 316.8 µl EB and 3.2 µl 2% Triton X-100) and incubating at room temperature for at least 1 h. Tubes were gently tapped halfway through elution. DNA was collected by transferring eluate with a P200 wide-bore pipette to a new 1.5 ml microcentrifuge tube. Some liquid and DNA remained on the Nanobind disk

after pipetting. The tube containing the Nanobind disk was spun in a centrifuge at 10,000*g* for 5 s, and any additional liquid that came off the disk was transferred to the eluate tube. This process was repeated if necessary until all DNA was removed. The samples were pipette mixed 5 times (approximately 10 s to aspirate and 10 s to dispense for each cycle) with a wide-bore P200 pipette to homogenize the sample. Samples were further allowed to rest at room temperature overnight to allow DNA to solubilize (disperse).

## Library preparation

### DNA tagmentation and FRA

Circulomics EB+ (EB buffer with 0.02% Triton X-100) was prepared, and 140.82 µl EB+ was aliquoted into a 1.5 ml Eppendorf DNA Lo-Bind tube. UHMW DNA (300 µl) from above was aliquoted into the same tube with a wide-bore P200 pipette. The mixture was slowly pipetted up and down 3 times with a wide-bore P200 pipette set to 150 µl. In a separate 1.5 ml Eppendorf DNA Lo-Bind tube, the following reagents were added in sequence: 144 µl of FRA dilution buffer, 9.18 µl of 1 M $MgCl_2$ and 6 µl FRA. The tube was tapped to mix and spun down using a microcentrifuge. The EB–Triton X-100–DNA mixture was added to the FRA dilution buffer–$MgCl_2$–FRA mixture with a wide-bore P200 pipette. This mixture was then pipette mixed 15–20 times with a wide-bore P1000 pipette set to 600 µl. The mixture appeared homogeneous when pipette mixing was finished. The tube was then incubated for 15 min at room temperature. The mixture was then pipette mixed 5 times with a wide-bore P1000 pipette set to 600 µl and incubated at room temperature for an additional 15 min. The mixture was incubated at 30 °C for 1 min, followed by 80 °C for 1 min and then held at 4 °C.

**FRA clean-up**

Clean-up used a Nanobind disk. A 5 mm Nanobind disk was added to the above-described reaction mixture followed by 300 µl of Circulomics buffer NAF10. The tube was gently tapped 10–20 times to mix. The mixture was placed on a platform rocker at 20 r.p.m. for 2 min at room temperature. A DNA ‘cloud’ was visible on the Nanobind disk. The tube was spun for 1–2 s using a benchtop microcentrifuge and placed on a magnetic rack. The binding solution was removed and discarded. The Nanobind disk was washed by adding 350 µl ONT long fragment buffer (LFB) and gently tapped 5 times to mix. The tube was spun for 1–2 s using a microcentrifuge and placed on a magnetic rack. The ONT LFB was removed and discarded. Care was taken to not pipette DNA attached to the Nanobind disk. This LFB wash was repeated. The tube was then briefly spun (microcentrifuge) to move the Nanobind disk to the bottom of the tube. DNA was eluted from the Nanobind disk by adding 125 µl of ONT EB to the tube. The tube was incubated for 30 min at room temperature then gently tapped 5 times (mixing) and incubated for an additional 30 min at room temperature. Fluid was slowly aspirated 4 times over the Nanobind disk before removing the eluate from the tube. The eluate was transferred to a new 1.5 ml Eppendorf DNA Lo-Bind tube using a wide-bore P200 pipette. The eluate was then pipette mixed 2 times with a wide-bore P200 pipette.

**Adapter attachment and rapid adaptor**

Rapid adaptor (RAP) was added to the DNA preparation. To 120 µl of eluate (from above), 3 µl of ONT RAP was added. The mixture was pipette mixed 8 times with a wide-bore P200 pipette. The mixture was then incubated for 15 min at room temperature and then again pipette mixed 8 times with a wide-bore P200 pipette.

**RAP reaction clean-up with Nanobind**

The final library clean-up step removes unligated adaptor. In brief, 120 µl Circulomics EB was added to 123 µl of the above-described RAP reaction

mixture. The mixture was slowly pipette mixed 3 times with a wide-bore P1000 pipette set to 240 µl. Each aspiration took about 10 s and each dispense took around 10 s. A 5 mm Nanobind disk was added to the reaction mixture followed by 120 µl Circulomics buffer NAF10. Mixing was accomplished by gentle tapping. The tube was incubated for 5 min at room temperature without agitation or rotation. The tube was gently tapped 5 times (each time for 2–3 times) during the 5 min of incubation. The tube was spun for 1–2 s using a microcentrifuge and placed on a magnetic rack. The binding solution was discarded. Next 350 µl of ONT LFB was added to the tube and mixed by gentle tapping 5 times. The tube was then spun for 1–2 s using a microcentrifuge and placed on a magnetic rack. The ONT LFB was removed and discarded. Next the Nanobind disk was washed by adding 350 µl ONT LFB. The tube was gently tapped 5 times to move LFB over the surface of the disk. The tube was then incubated at room temperature for 5 min. The tube was then spun for 1–2 s using a microcentrifuge and placed on a magnetic rack. The ONT LFB was removed and discarded. The tube was briefly spun using a microcentrifuge to move the Nanobind disk to the bottom of the tube. To elute DNA from the Nanobind disk, 126 µl ONT EB was added to the tube. The tube was incubated for 30 min at room temperature, then gently tapped 5–10 times and incubated for an additional 1–2 h at room temperature. The eluate was then transferred to a new 1.5 ml Eppendorf DNA Lo-Bind tube using a wide-bore P200 pipette using the same technique described above for passing the eluate over the Nanobind disk before removing the eluate from the tube. The mixture was then pipette mixed 2–3 times with a wide-bore P200 pipette. The library was stored overnight at 4 °C before sequencing to facilitate maximal dissolution of DNA.

**Flow cell loading and sequencing**

ONT sequencing buffer (SQB) (68 µl) was added to 82 µl of the eluate from above. The mixture was pipette mixed 4 times with a wide-bore P200 pipette set to 150 µl. Each aspiration of 150 µl took 10–20 s, and each dispense of 150 µl took 10–20 s. Samples were then incubated at room temperature for 10 min. Next the samples were pipette mixed 8 times with a wide-bore P200 pipette set to 150 µl. Before loading the library, the flow cell was primed with flush buffer/flush tether mixture per ONT directions.

The library was then added to the flow cell. The mixture was viscous but loaded smoothly in about 1 min. Some samples took 2 min maximum to load. The sequencing run had a re-mux time set for every 6 h. Base calling was performed using Guppy (v.4.0.11), with default parameters and the high-accuracy PromethION model (dna_r9.4.1_450bps_hac_prom.cfg).

### Dovetail Omni-C

We prepared Omni-C libraries from each cell line using a Dovetail Omni-C kit (Dovetail Genomics) with the following modifications as. First, we aliquoted 1 million cells for fixation with formaldehyde and DSG. We digested chromatin with DNAse I until DNA fragments of a desired length were obtained. Per the protocol, we performed end repair on the chromatin, followed by ligation of a biotinylated bridge oligonucleotide, followed by ligation of free chromatin ends. We reversed crosslinks and purified proximity ligated DNA. We converted the DNA into an Illumina sequencing library using a NEB Ultra II library preparation kit (NEB) with a Y-adaptor. We enriched for ligation products using streptavidin bead capture on the final library. Each capture reaction was then split into two replicates before the final PCR enrichment step to preserve complexity. All libraries were uniquely dual indexed and sequenced on an Illumina Novaseq Platform with read lengths of 2 × 150 bp.

## Phased assembly pipeline

We describe the main automated and manual steps taken before, during and after assembly. A combined set of workflow description language (WDL) formatted assembly workflows is available from Dockstore that captures each of the steps for filtering adapter-contained reads and running Hifiasm (https://dockstore.org/organizations/HumanPangenome/collections/Hifiasm). All assemblies were generated using this workflow, running on AnVIL[71]. Cleaning assemblies and fixing some structural issues were performed through a combination of automated workflows and manual curation as described below. Manual curation was performed using Jupyter notebooks available at GitHub (https://github.com/human-pangenomics/hpp_production_workflows/tree/master/assembly/y1-notebooks).

**Filtering adapter-contained reads and running Hifiasm**

Before producing the assemblies, we detected and removed the reads containing PacBio adapters using a bash script from the HiFiAdapterFilt repository[72] (commit 64d1c7b). This script first creates a database of the PacBio adapter sequences, as illustrated below:

>gnl|uv|NGB00972.1:1-45 Pacific Biosciences Blunt Adapter
ATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAT
>gnl|uv|NGB00973.1:1-35 Pacific Biosciences C2 Primer
AAAAAAAAAAAAAAAAAATTAACGGAGGAGGAGGA

It then runs blastn with tuned parameters to detect adapter-containing reads as follows:

blastn -db ${DATABASE} -query ${HIFI_FASTA} -task blastn -reward 1 -penalty -5 -gapopen 3 -gapextend 3 -dust no -soft_masking true -evalue 700 -searchsp 1750000000000 -outfmt

For 43 samples (out of 47), we removed less than 0.15% of the reads; all of the 29 HPRC-selected samples are among these 43 samples, indicating the low level of adapter contamination in the HiFi data produced by the HPRC. HG005, which is one of the 18 additional samples, had the highest adaptor contamination percentage, at about 1%. (Supplementary Fig. 39)

The removed reads were then aligned to the T2T-CHM13 (v.2.0) reference to ensure that there was no chromosomal or locus-specific bias in the filtering process. Supplementary Fig. 40 shows a snapshot of the IGV browser[73] illustrating the coverage of the adapter-containing reads along the genome. The locations of the reads were almost evenly distributed along the genome and, excluding centromeres, we barely found any region covered with more than two adapter-containing reads, even in HG005, which had the highest contamination percentage.

The trio-binning mode of Hifiasm needs haplotype-specific *k*-mers for trio phasing the assembly graph. To generate these *k*-mers, we used parental Illumina short reads for the 47 HPRC samples, which are publicly available

from the 1KG dataset[19]. For each parental short-read sample, we used yak count (v.0.1)[74] to generate the *k*-mer hash tables, running it once for each of the paternal and maternal read sets:

yak count -k31 -b37 -o pat.yak paternal.fq.gz yak count -k31 -b37 -o mat.yak maternal.fq.gz

The adapter-filtered HiFi reads along with the parental *k*-mer tables were then given to Trio-Hifiasm (v.0.14) to produce haplotype-resolved assembly graphs. Only the sample HG002 was re-assembled with Trio-Hifiasm (v.0.14.1), which is explained in more detail in the next subsection.

hifiasm -o ${SAMPLE_NAME} -t 48 -1 pat.yak -2 mat.yak hifi.fq.gz

Hifiasm produces one graph per haplotype in GFA format. Each haplotype-specific GFA file was then converted to FASTA format using Gfatools[75]. The assemblies produced by Trio-Hifiasm (v.0.14) are released under v.2 after performing the three cleaning steps described at the end of this section.

**Manually fixing issues**

We used paftools.js asmgene, from the minimap2 repository (https://github.com/lh3/minimap2/tree/master/misc)[76] to count the number of apparent gene duplications for each of the assemblies produced by Trio-Hifiasm (v.0.14). Asmgene does not distinguish between true duplicates and errors. Looking at its results, we were able to find the duplication trend and detect any outlier. This assessment acted as a proxy for detecting high-level duplication errors. We used the Ensembl v.99 cDNA sequences[77] as the input gene set for running asmgene.

# Aligning genes to GRCh38 and each Hifiasm haploid assembly minimap2 -cx splice:hq hs38.fa cdna.fa > hs38.paf minimap2 -cx splice:hq ${pat/mat}.fa cdna.fa > ${pat/mat}.paf # Detecting gene duplications paftools.js asmgene -a hs38.paf ${pat/mat}.paf

Three samples were detected as outliers in terms of the number of gene duplications. To identify the cause of this issue, we aligned back the HiFi

reads to those assemblies and checked the depth of coverages and mapping qualities. It showed that the samples HG01358, HG01123 and HG002 contained false duplications in at least one haplotype of length around 55 Mb (in h1tg000058l contig), about 14 Mb (h1tg000013l) and around 70 Mb (h2tg000045l), respectively. In the assembly graphs of HG01358 and HG01123, the duplicated HiFi reads that appeared multiple times were used as anchors to manually determine the exact boundaries of the duplicated regions in the contigs. These two contigs were then manually fixed by breaking the contigs at the duplication start and end points and discarding the duplicated sequence from the assembly. In detail, for HG01123 for h1tg000013l, we discarded the interval [94439457, contig end]. For HG01358 for h1tg000058l, we kept the interval [0, 95732608), renaming the contig to h1tg000058l_1, we discarded the interval [95732608, 150395342) and kept the interval [150395342, contig end], renaming it to h1tg000058l_2. To address the false duplication in HG002 we re-assembled it using a newer version of Trio-Hifiasm (v.0.14.1), which was reported not to have this problem.

We also evaluated the phasing accuracy of the assemblies by using yak trioeval (see below). We detected a single large misjoin in a maternal contig of the HG02080 assembly. It contained an approximately 22-Mb-long paternal block in the middle of the contig, which resulted in two switch errors at the edges of this block. This block was manually discarded from the assembly and the contig was broken into two smaller ones. In detail, in HG02080 for the h2tg000053l contig, we kept the interval [0, 41506503), renaming it to h2tg000053l_1, we discarded the interval [41506503, 63683095) and kept the interval [63683095, contig end], renaming it to h2tg000053l_2.

We finally searched for interchromosomal misjoins using the Minigraph pangenome (see below for construction details). An ‘interchromosomal misjoin’ was defined by a chimeric Minigraph alignment (see below) consisting of ≥1 Mb subalignments on different chromosomes.

**Cleaning steps**

To clean the raw assemblies, we performed three additional steps: masking the remaining HiFi adapters, dropping the contigs that were contaminated in their entirety and removing any redundant mitochondrial contigs.

In the first cleaning step, the sequence of the PacBio SMRTbell adapter was aligned to each assembly using minimap2 with the parameters -cxsr -f5000 -N2000 --secondary=yes. We extracted only the hits with less than or equal to 2 mismatches and which were longer than 42 nt. In addition, eukaryotic adapters in each assembly were identified by VecScreen[78]. The combined minimap2 and VecScreen adaptor hits (when present) were hard-masked in the assemblies using a WDL of the bedtools maskfasta command (https://dockstore.org/workflows/github.com/human-pangenomics/hpp_production_workflows/MaskAssembly:master?tab=info).

bedtools maskfasta \ -fi ${inputFastaFN} \ -bed ~{adapterBed} \ -fo ~{outputFasta}

In the second cleaning step, we used VecScreen to detect mitochondrial contigs and contigs consisting of nonhuman sequences from other organisms, such as bacteria, viruses and fungi. These contigs were then dropped from the assemblies using a WDLized version of samtools faidx. It is worth noting that the contigs with nuclear mitochondrial DNAs within them were not dropped.

samtools faidx \ $inputFastaFN \ `cat contigsToKeep.txt` | gzip \ > ~{outputFasta}

In the last cleaning step, we selected one contig as the best mitochondrial contig per diploid assembly. To do this, selection we aligned the sequence of the mitochondrial DNA (with the RefSeq identifier of NC_012920.1) to each diploid assembly using minimap2 with the parameters -cx asm5 --cs. Then we selected one contig with the highest mapping score and the lowest number of mismatches as the best mitochondrial contig (we selected one randomly if multiple best contigs existed). This contig was then rotated and flipped (if necessary) to match the start and orientation of NC_012920.1.fa and then added to the maternal assembly of the corresponding sample. Only the HG01071 sample did not produce any identifiable mitochondrial contig.

Masked and cleaned mitochondrial assemblies were then accessioned to GenBank, where they underwent another round of adapter masking and removal of contamination, which was mostly Epstein–Barr virus used to generate the LCLs. The final cleaned assemblies in GenBank were downloaded, and the contig identifiers were pre-pended with the sample name and haplotype integer (where 1 = paternal and 2 = maternal). For example, a contig assigned the name JAGYVH010000025 in sample HG02257's maternal assembly was renamed to be HG02257#2#JAGYVH010000025. The renamed assemblies were then released to our Amazon Simple Storage Service (S3) and Google Cloud Platform (GCP) buckets. In the process of downloading from GenBank, three of the assemblies (HG00733 paternal, HG02630 paternal and NA21309 maternal) had their downloads prematurely stopped, which resulted in missing sequences. Notably, NA21309 is missing its mitochondrial contig. Details can be found on the HPRC's year 1 assembly GitHub repository (https://github.com/human-pangenomics/HPP_Year1_Assemblies). The assemblies held at the International Nucleotide Sequence Database Collaboration are not truncated, but the truncated copies were retained in S3 and GCP as they were used in the construction of the pangenomes.

After submission to GenBank, the assemblies were aligned against CHM13 using Winnowmap, and multiple contigs were found to be unmapped. These contigs were subjected to BLAST and found to be almost exclusively Epstein–Barr virus sequence. GenBank confirmed (personal communication) that these unmapped contigs should have been dropped as contamination, but since the genomes were already in active use, they elected not to remove them at this time. A list of the contigs that should have been dropped can be found on the year 1 assembly GitHub repository (https://github.com/human-pangenomics/HPP_Year1_Assemblies/blob/main/genbank_changes/y1_genbank_remaining_potential_contamination.txt).

## Assembly assessment pipeline

Several steps in assembly assessment were managed through a StandardQC workflow written using WDL, run on AnVIL, and available at Dockstore

(https://dockstore.org/workflows/github.com/human-pangenomics/hpp_production_workflows/StandardQC). Individual tools within the workflow were run in Docker containers with specific tool versions installed for consistency and reproducibility. Details are available within the Dockstore-deposited workflow. The StandardQC workflow takes short-read data for parental and child samples, the two assembly haplotypes, and it produces an analysis over various quality metrics produced by the tools described below. For each task, the workflow produced a small human-readable summary file, which is also easy to parse for summarizing steps, as well as the full output from the tool for manual inspection. Specific tool invocations can be determined from the deposited workflow and are described in the subsequent sections.

**Measuring interchromosomal joins**

Contigs were aligned to CHM13 (v.2.0) with Minigraph (v.0.18) and processed with the following command line:

minigraph -cxasm chm13v2.0.fa contigs.fa | paftools.js misjoin –

The ‘misjoin’ command reports an interchromosomal join if a contig has two ≥1 Mb alignments to two different chromosomes.

**Assembly contiguity assessment**

Assembly contiguity was assessed for each haplotype using QUAST[79]. These statistics included total sequence assembled, total assembled contigs and contig NG50 (assuming a genome size of 3.1 Gb). All reference-based analyses were skipped.

QUAST was invoked with the following command:

python /opt/quast/quast-5.0.2/quast-lg.py -t 16 -o <sample>.quast --large --est-ref-size 3100000000 --no-icarus

**Assembly QV assessment**

Assembly QV was determined using two separate *k*-mer-based tools. The first is Yak[20]. Yak's QV estimation happens separately on each haplotype. The *k*-mer databases for Yak were generated using the following command:

```
yak count -t16 -b37 -o <sample>.yak <(cat <read_files>) <(cat
<read_files>)
```

QV estimation using Yak was generated with the following command:

```
yak qv -t 32 -p -K 3.2g -l 100k <sample>.yak
<sample_assembly_haplotype> > <sample>.<haplotype>.yak.qv.txt
```

Assembly QV was also determined using Meryl and Merqury[80]. Meryl generates *k*-mer databases and Merqury determines haplotype QV jointly with both haplotypes.

The *k*-mer databases with Meryl were generated with the following commands. Databases were generated separately for each read file using meryl count and merge with meryl union-sum. Parental-specific *k*-mers (hap-mers) were generated using merylu hapmer.

```
meryl k=21 threads=64 memory=32 count output <sample>.meryl
<read_file> meryl union-sum output <sample>.meryl
<sample_read_meryl_files> bash hampers.sh maternal.meryl paternal.meryl
sample.meryl
```

QV estimation using Merqury was generated with the following command:

```
merqury.sh sample.meryl maternal.meryl paternal.meryl
<maternal_haplotype> <paternal_haplotype> <sample>.merqury
```

**GIAB-based assembly quality assessment**

As a complementary and stratified assessment of assembly quality, we used the GIAB assembly benchmarking pipeline to compare assembly-based variant calls to GIAB's small variant benchmarks (v.4.2.1) for two GIAB samples assembled in this work: HG002 and HG005. We evaluated the HG002 and HG005 HPRC assemblies aligned to GRCh38. Variants were

called from assemblies using Dipcall (v.0.3) (using mimimap2 (v.2.2.4))[38]. We used -z200000,10000 parameter to improve alignment contiguity, as it was previously shown to improve variant recall in regions with dense variation, such as the major histocompatibility complex[81]. Small variant evaluation was performed using hap.py (v.3.15)[82], benchmarking against v.4.2.1 of high-confidence SNP, small indel and homozygous reference calls for the GIAB samples HG002 and HG005. Comparisons were performed with and without restriction to the associated dipcall region file (dip.bed) to assess recall within and outside assembled regions. For better comparisons of complex variants, hap.py was run using vcfeval[83]. Variant calls were stratified using GIAB stratifications (v.3.0)[84], stratifying true-positive, false-positive and false-negative variant calls in challenging and targeted regions of the genome.

### Trio-based assembly phasing assessment

Assembly phasing was assessed using Yak and described using two statistics: switch error and Hamming error rates. Switch error describes the number of times two adjacent phased variants incorrectly switch between maternal and paternal haplotypes. Hamming error rate relates to the total number of misphased variants per assembled contig. Yak generates phasing statistics separately for each haplotype using parental *k*-mers gathered from Illumina short-read sequencing of the parents.

Yak generates *k*-mer databases for the sample and both parental haplotypes (as described above). We used Yak to generate phasing metrics with the following command:

yak trioeval -t 32 paternal.yak maternal.yak <haplotype_assembly> >
<sample>.<haplotype>.yak_phasing.txt

### Hi-C-based assembly phasing assessment

An alternative approach for phasing evaluation is to use Hi-C reads that do not require trio information. We computed the switch error rate for local phasing evaluation and the Hamming error rate for global phasing

evaluation. We implemented an efficient *k*-mer-based method in pstools (v.0.1)[24] and used maximum Hi-C read support to detect switch errors on heterozygous positions. In this procedure, we first identified heterozygous *k*-mers (hets) from phased assemblies using 31-mers. Then we mapped Hi-C reads to the assemblies using these 31-mers. If there were >5 reads that supported a switch between consecutive hets in assemblies, we considered a haplotype switch. For each het pair, we noted whether Hi-C reads supported or did not support the phase. We considered a switch error when a het site had a phase switched support relative to that of the previous heterozygous site. The switch error rate is the number of local switches divided by the number of heterozygous sites. We performed this operation for the entire contig over all contigs for switch calculations. For the Hamming error calculations, we considered Hamming distance on the entire contig level divided by the number of heterozygous sites. This measure gives a global view of phasing errors and implicitly penalizes any long switches in contigs.

## Assembly read-based evaluation of Flagger

The following describes the generation and cleaning of the HiFi alignments to the HPRC assemblies and running Flagger (v.0.1), a read-based pipeline for evaluating diploid/dual assemblies. All the WDL-based workflows for running these steps are deposited in the Dockstore collection (https://dockstore.org/organizations/HumanPangenome/collections/Flagger-Secphase).

### Preparing the HiFi alignments

We aligned back the HiFi reads of each sample to its diploid assembly. The alignments were produced with winnowmap (v.2.03) using the following commands:

# making the k-mer table with meryl meryl count k=15 output merylDB asm.fa meryl print greater-than distinct=0.9998 merylDB > repetitive_k15.txt # alignment with winnowmap winnowmap -W repetitive_k15.txt -ax map-pb -Y -L --eqx --cs -I8g <(cat pat_asm.fa mat_asm.fa) reads.fq.gz | samtools view -hb > read_alignment.bam

For all samples, we used the full HiFi read sets mentioned in Supplementary Table 1, except HG002, for which we downsampled the read set to 35×.

To exclude unreliable alignments, we removed all chimeric alignments and alignments shorter than 2 kb or with a gap-compressed mismatch ratios higher than 1%. As the assembly is diploid and the reads aligned to the homozygous regions are expected to have low mapping qualities, we did not filter alignments on the basis of their mapping qualities. In Supplementary Fig. 41, we plot the histograms of mapping qualities and the distributions of alignment identities for one sample, HG00438, as an example. The statistics of three sets of alignments were plotted: the alignments to the diploid assembly and to each haploid assembly (maternal and paternal) separately. It indicates that the reads have higher identities when the diploid assembly is used as reference but about 20% more reads have mapping qualities lower than 10.

Generally, in highly homozygous regions, the aligner may not be able to select the correct haplotype as the primary alignment because of either read errors or misassemblies. To detect these cases, we searched for secondary alignments for which the scores were almost as high as the primary alignment of the same read. For each such read, we made a pseudo-multiple alignment of the read sequence and the assembly blocks captured by all secondary and primary alignments. Using this alignment, we searched for the read bases that were mismatched in at least one alignment but not all alignments. We called such bases single nucleotide markers. For each alignment, we calculated a consistency score by considering only the single nucleotide markers and taking the summation of their base qualities with a negative sign. We then sorted the alignments (regardless of being primary or secondary) based on this score. If the best alignment was a secondary alignment, we assigned the primary tag to this alignment and removed the other alignments. The percentage of the total reads with swapped alignments ranged from 0.03% (HG03453) to 0.44% (HG005) across 47 HPRC samples. This result shows that only a small percentage of the reads needed to be relocalized using this method. This step was performed through the Secphase (v.0.1) workflow, which is available in the Dockstore collection

(https://dockstore.org/organizations/HumanPangenome/collections/Flagger-Secphase).

By calling variants, it is possible to detect the regions that either need polishing (that is, are errors) or that have alignments from the wrong haplotype because of mismappings. We used DeepVariant (v.1.3.0) with the parameter --model_type="PACBIO" to call variants on these alignments. The variants were then filtered to include only the biallelic SNPs with a variant frequency higher than 0.3 and genotype quality higher than 10.

bcftools view -Ov -f PASS -m2 -M2 -v snps -e 'FORMAT/VAF < 0.3 || FORMAT/GQ < 10' ${OUTPUT_VCF} > ${SNPS_VCF}

Having the biallelic SNPs, we found the alignments with alternative alleles and removed them from the bam file. For this step, we implemented and used the program filter_alt_reads, running the following command:

filter_alt_reads -i ${INPUT_BAM} -o ${ALT_FILTERED_BAM} -f ${ALT_BAM} -v ${SNPS_VCF}

**Running the evaluation pipeline**

To assess the read mappings resulting from our diploid alignment process, we used the following five steps, which are combined into a pipeline that we refer to as Flagger. Flagger essentially fits a mixture model to successive coverage blocks of the read-to-diploid assembly alignment and then classifies each block to a category predicting the accuracy of the assembly at that location.

**Step 1: calculating depth of coverage**

First, after producing and cleaning the HiFi alignments, we calculated the depth of coverage for each assembly base by samtools depth -aa (the -aa option allows outputting the bases with zero coverage):

samtools depth -aa -Q 0 read_alignment.bam > read_alignment.depth

The output of samtools depth was then converted into a more efficient format with the .cov suffix. This format is implemented specifically for Flagger and is more efficient, as the consecutive bases with the same coverage take only one line. We implemented a program called depth2cov for converting the output of samtools depth to the .cov format.

depth2cov -d read_alignment.depth -f asm.fa.fai -o read_alignment.cov

**Step 2: fitting the mixture model**

In the second step, the frequencies of coverages were calculated using cov2counts. The output file with the .count suffix is a two-column tab-delimited file: the first column shows coverages and the second column shows the frequencies of those coverages.

cov2counts -i read_alignment.cov -o read_alignment.counts

The python script fit_gmm.py takes a file .counts suffix, fits a Gaussian mixture model and finds the best parameters through expectation-maximization. This mixture model consists of four main components and each component represents a specific type of region:

1. (1)

   Erroneous component, which is modelled by a Poisson distribution. To avoid overfitting, this mode only uses the coverages below 10 so its mean is limited to be between 0 and 10. It represents the regions with very low read support.

2. (2)

   (Falsely) duplicated component, which is modelled by a Gaussian distribution, the mean of which is constrained to be half of mean of the haploid component. It should mainly represent the falsely duplicated regions.

3. (3)

Haploid component, which is modelled by a Gaussian distribution. It represents blocks with the coverages that we expect for the blocks of an error-free assembly.

4. (4)

   Collapsed component, which is actually a set of components each of which follows a Gaussian distribution, the mean of which is a multiple of the mean of the haploid component. It represents regions that have additional copies present in the underlying genome that have been ‘collapsed’ into a single copy.

It was noted that the model components may change for different regions owing to regional coverage differences and that the resulting systematic differences affect the accuracy of the partitioning process. To make the coverage thresholds more sensitive to the local patterns, the diploid assembly was split into windows of length (5–10 Mb) and a distinct model was fit for each window. Before fitting, we split the whole-genome coverage file produced in the first step into multiple coverage files for each window. We implemented and ran split_cov_by_window for splitting:

```
split_cov_by_window -c read_alignment.cov -f asm.fa.fai -s 5000000 -p ${OUTPUT_PREFIX}
```

This produced a list of coverage files, each of which ends with ${CONTIG_NAME}_${WINDOW_START}_${WINDOW_END}.cov

We then repeated the above-described steps for each resulting coverage file.

One important observation is that for short contigs, the coverage distribution is generally too noisy to satisfactorily fit the mixture model. To address this issue, we performed the window-specific coverage analysis only for the contigs longer than 5 Mb. For the shorter contigs, we used the results of the whole-genome analysis.

**Step 3: extracting blocks of each component**

Using the fitted model, we assigned each coverage value to one of the four components (erroneous, duplicated, haploid and collapsed). To do so for each coverage value, we picked the component with the highest probability. For example, the coverage value 0 is frequently assigned to the erroneous component. In Supplementary Fig. 42, the coverage intervals are coloured based on their assigned component.

**Step 4: incorporating coverage biases in HSats**

According to an article describing a complete human genome[3], there are some satellite arrays (for example, HSat1, HSat2 and HSat3) for which the HiFi coverage drops or increases systematically owing to biases in sample preparation and sequencing. Such platform-specific biases mislead the pipeline. As a result, the falsely duplicated component may contain a mixture of falsely duplicated and coverage-biased blocks. Similar effects occur for the collapsed component.

To incorporate such coverage biases and to correct the results in the corresponding regions, we first found the regions of each haploid assembly for which a coverage bias is expected. To find such regions, we lifted over the CHM13 HSat1, HSat2 and HSat3 annotation to each assembly by aligning the assembly contigs to the reference T2T-CHM13 (v.1.1) and GRCh38 (chromosome Y) and projecting the HSat coordinates back to the assembly (using python script project_blocks.py). Then we ran fit_gmm.py to fit a mixture model for the blocks assigned to each HSat type and adjusted the parameter --coverage, the starting point of the expectation-maximization process, based on the expected coverage in the corresponding HSat. For HSat1, HSat2 and HSat3 we set --coverage to 0.75, 1.25 and 1.25 times the average sequencing coverage, respectively. Finally, we decomposed each HSat based on the inferred coverage thresholds and replaced the previous assigned component by the new one.

**Step 5: using high-quality alignments to correct spurious flags**

In some cases, the duplicated component was mixed up with the haploid one. It usually happens when the coverage in the haploid component drops

systematically or the majority of a long contig is falsely duplicated. To address this issue, we used another indicator of a false duplication, which is the accumulation of alignments with very low mapping quality (MAPQ). We produced another coverage file using only the alignments with MAPQ > 20. Whenever we found a region flagged as duplicated with more than five high-quality alignments, we changed the flag to haploid.

After the correction made in step 5, we merged blocks from each component closer than 1,000, and the overlap of any two components after merging was flagged as 'unknown' to show that this block could not be properly assigned. The BED files produced by Flagger are available in the HPRC S3 bucket (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/e9ad8022-1b30-11ec-ab04-0a13c5208311--COVERAGE_ANALYSIS_Y1_GENBANK/FLAGGER/APR_08_2022/FINAL_HIFI_BASED/FLAGGER_HIFI_ASM_SIMPLIFIED_BEDS/).

**Assessing T2T-CHM13 using Flagger**

To estimate the false-positive rate of Flagger, we applied it to the T2T-CHM13 (v.1.1) reference. The direct output of Flagger showed that about 12.77 Mb (around 0.41%) of the T2T-CHM13 reference assembly was flagged as potentially unreliable. The HPRC assemblies were almost free of rDNA arrays, but there were modelled sequences for rDNA arrays in the T2T-CHM13 (v.1.1) reference. These arrays were flagged as falsely duplicated in their entirety, which indicated that Flagger with HiFi reads may not be able to correctly evaluate rDNA arrays. Therefore, to make a fair comparison, we excluded rDNA arrays (about 9.92 Mb in total) from the reference evaluation, which decreased the number of unreliable bases to 5.58 Mb (around 0.18%). We additionally identified about 2.76 Mb of a region beside chromosome 1–HSat2 that was mis-flagged as collapsed. This mis-flagging was the impact of the systematic coverage rise on the neighbouring HSat2 that altered the fitted mixture model. By manually fixing this mis-flagging, we had about 2.82 Mb (0.09%) of unreliable blocks in T2T-CHM13 (v.1.1). This number is about 9.3 times lower than the average for the HPRC assemblies. These unreliable blocks are mainly a combination of 'Unknown' blocks, which could not be properly assigned

and the regions with HiFi-specific coverage drops. The results of this analysis are available in the HPRC S3 bucket (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=submissions/e9ad8022-1b30-11ec-ab04-0a13c5208311--COVERAGE_ANALYSIS_Y1_GENBANK/FLAGGER/APR_08_2022/FINAL_HIFI_BASED/T2T-CHM13/).

## Repeat masking

Repeat masking on each assembly was iteratively performed using RepeatMasker (v.4.1.2-p1). The first step masked used the default human repeat library, and the second step used a repeat library augmented by CHM13 satellite DNA sequences on the original assemblies after hard masking the initial repeat masked DNA. The augmented repeat library (final_consensi_gap_nohsat_teucer.embl.txt) is available at Zenodo (https://doi.org/10.5281/zenodo.5537107), and a parallelized repeat masking pipeline (RepeatMaskGenome.snakefile) is available at GitHub (https://github.com/chaissonlab/segdupannotation). The union of the two steps generated the complete repeat masking.

## SD annotation

SDs were annotated using sedef[85] after masking repeats in each assembly. Repeats annotated with more than 20 copies corresponded to unannotated mobile elements and were excluded from the analysis. The pipeline for annotating SDs is available at GitHub (https://github.com/ChaissonLab/SegDupAnnotation/releases/tag/vHPRC).

## SD reliability

The reliable and unreliable regions for all haplotype assemblies were aligned to T2T-CHM13 (v.2.0) and then subdivided into 5 kb windows and intersected with the SD annotations for T2T-CHM13. SD annotations were unavailable for chromosome Y on T2T-CHM13 (v.2.0) at the time of analysis. Furthermore, the chromosome Y added to T2T-CHM13 is from the HPRC HG002 sample; therefore chromosome Y was excluded. For each class of unreliable region (unknown, erroneous, duplicated, collapsed and

haploid), we calculated the average number of base pairs overlapping SDs across the haplotype assemblies and annotated each 5 kb window with the most representative overlapping SD (the SD with the highest product of identity and length). Then using the most representative SD, we calculated the average length and identity of SDs overlapping each class of unreliable region for all the 94 haplotypes and compared the length of identity of SDs that overlapped the different types of errors in the assembly. The code for this analysis is available on GitHub (https://gist.github.com/mrvollger/3bdd2d34f312932c12917a4379a55973).

## Ensembl mapping pipeline for annotation

To create high-confidence annotations, a new Ensembl annotation pipeline was developed. The pipeline clusters and maps spatially proximal genes in parallel (to help avoid issues with individually mapping near identical paralogues) and attempts to resolve inconsistent mappings by both considering the synteny of the gene neighbourhood in relation to the GRCh38 annotation and the identity and coverage of the underlying mappings.

A reference gene set was created from a subset of the GENCODE (v.38) genes[29], which was mapped to the HPRC assemblies through a two-pass alignment process. This excluded readthrough genes and genes on patches or haplotypes, and only included one copy of the genes on the X/Y PAR region (only one copy, chromosome X, is modelled in the Ensembl representation of the *PAR* genes).

First, to minimize the difficulty of mapping near identical paralogues, a jumping window of 100 kb in length was used to identify clusters of genes to map in parallel (Supplementary Fig. 43). The initial window was positioned at the start of the most 5′ gene for each chromosome in the GRCh38 reference and extended 100 kb from the start of the gene. Any genes fully or partially overlapping the window were then included in the cluster. The next 3′ gene that did not overlap the previous window was then identified, and a new window was created and the process repeated. This resulted in both clustered genes and non-clustered genes (genes were considered not clustered when there was only one gene within the window).

The regions to map were then identified on the basis of the start of the most 5′ gene and the end of the most 3′ gene in each cluster (or simply the 5′ and 3′ end of the gene in the case of non-clustered genes).

For each region defined in the previous step, anchor points were then selected to help map the region on the target genome. Two 10 kb anchor points were created 5 kb from the 5′ and 3′ edge of the region, and a central 10 kb anchor was created around the midpoint of the region in the GRCh38 genome. The sequences of these anchors were them mapped against the target genome using minimap2 (ref. [86]) with the following command:

```
minimap2 --cs --secondary=yes -N 10 -x map-ont [genome_index] [anchor_file] > [alignment_file]
```

The resulting hits were examined to determine high-confidence regions in the target genome. High-confidence regions were ones in which all three anchors were on the same top-level sequence, in colinear order, with ≥99% sequence identity and ≥50% hit coverage, and with a similar distance between the anchors when compared with the reference genome. If no suitable candidate region was found with all three anchors, pairs of mapped anchors were then assessed in a similar manner.

The sequence-selected region or regions were then retrieved and aligned against the corresponding GRCh38 region using MAFFT. For each gene, the corresponding exons were retrieved and the coordinates were projected through the alignment of the two regions. Transcripts were then reconstructed from the projected exons. For each transcript, the coverage and identity when aligned to the parent transcript from GRCh38 were calculated.

If the resulting transcript had either a coverage of <98% or an identity of <99%, the parent transcripts were aligned to the target region using minimap2 in splice-aware mode, with the high-quality setting for Iso-Seq/cDNA transcripts enabled. The maximum intron size was set to 100 kb by default. For transcripts with reference introns larger than 100 kb, the maximum intron size was scaled and set as 1.5 times the length of the longest intron (to allow some variability):

```
minimap2 --cs --secondary=no -G [max_intron_size] -ax splice:hq -u b [expected_target_region] [transcript_sequences] > [sam_file]
```

For each transcript that mapped to the target genome, the quality of the mapping was assessed on the basis of aligning the original reference sequence with the newly identified target sequence. Again, if the coverage or identity of the aligned sequence was <98% or <99%, respectively, the reference transcript sequence was re-aligned to the target region, this time using Exonerate[87]. Exonerate, although slower than minimap2, has the ability to handle very small exons and can incorporate CDS data to preserve the CDS (introducing pseudo-introns as needed). The following command was used:

```
exonerate -options --model cdna2genome --forwardcoordinates FALSE --softmasktarget TRUE --exhaustive FALSE --score 500 --saturatethreshold 100 --dnawordlen 15 --codonwordlen 15 --dnahspthreshold 60 --bestn 1 --maxintron [max_intron_size] -coverage_by_aligned 1 --querytype dna --targettype [target_type] --query [query_file] --target [target_file] --annotation [annotation_file] > [output_file]
```

When more than one approach was used to model the transcript, the mapping with the highest combined identity and coverage was selected.

For genes not mapped through the initial regional anchors, a second approach was used. The expected location of the gene was located using high-confidence genes mapped during the first phase. High-confidence mappings were those for which there was a single mapped copy of the gene, all the transcripts had mapping scores of 99% coverage and identity on average and the gene also had a similar gene neighbourhood to the neighbourhood in the reference (at least 80% of the of the same genes in common for the 100 closest neighbouring genes in the reference). After this step, the entire genome region underlying the missing gene, including a 5 kb flanking sequence, was mapped against the target genome using minimap2:

```
minimap2 --cs --secondary=yes -x map-ont [genome_index] [gene_genomic_sequence] > [alignment_file]
```

更多杂志请加vx
几百种杂志持续更新

The resulting hits were then filtered on the basis of overlap with the expected region that the missing gene should lie in. If there was no expected region calculated (cases in which no pair of high-confidence genes could be found to define the 5′ and 3′ boundaries of the expected location of the missing gene, for example, at the edge of a scaffold) or no hit overlapping the expected region was found, the top reported hit was used providing it passed an identity cut-off of 99%. The selected hit or hits were then extended on the basis of how much of the original reference gene they covered to ensure that minor local variants between the reference and target regions did not lead to the target region being truncated. Once extended, the remaining hits were then clustered on the basis of genomic overlap and merged into unique regions. The missing genes were then attempted to be mapped to these regions using an identical process as described above for the initial mappings, involving MAFFT, minimap2 and Exonerate.

To minimize the occurrences of mis-mapped paralogues, each gene was checked for exon overlap in both the target and the reference. If the overlapping genes were not identical at a locus between the reference and the target, then a conflict was identified. For each gene present, filtering was done to reduce or remove the conflict based on a number of factors, including whether the genes were in the expected location, whether the genes were high-confidence mappings, the average percent identity and coverage of the transcript for the genes and the neighbourhood score. When it was not possible to resolve a conflict between two genes, both were kept. This concluded the primary mapping process.

After this process, potential recent duplications were identified. To search for recent duplications, the canonical transcript of each gene (the longest transcript in the case of noncoding genes, or the transcript with the longest translation followed by the longest overall sequence for protein-coding genes) was selected and aligned across the genome using minimap2 in a splice-aware manner:

```
minimap2 --cs --secondary=no -G [max_intron_size] -ax splice:hq -u b
[genome_index] [input_file] > [sam_file]
```

Mappings that had exon overlap with existing annotations from the primary mapping process on the target genome were removed. For new mappings

that did not overlap existing annotations, the quality of the alignment was then assessed by aligning the mapped transcript sequence to the corresponding reference transcript to calculate the coverage and per cent identity of the mapping. Different coverage and per cent identity cut-offs were used for these mappings on the basis of the type of transcript mapped. Protein-coding and small noncoding transcripts used a coverage and identity cut-off of 95%, whereas long noncoding transcripts used a coverage and identity cut-off of 90%. Pseudogene transcripts had a lower coverage cut-off of 80%, but the same identity cut-off of 90% as long noncoding transcripts.

When looking for new paralogues, for cases in which multiple canonical transcripts mapped to a locus, a single representative transcript was selected. This was based on the following hierarchy of gene biotype groups: coding, long noncoding, pseudogene, small noncoding, and miscellaneous or undefined.

If there were multiple transcripts for the highest represented group, the transcript with the longest sequence was selected as the representative.

## Gene annotation quality analysis

### Frameshifts

For the Ensembl and CAT gene annotation sets, we identified the locations of frameshifting indels by iterating over the coding sequence of each transcript and looking for any gaps in the alignment. If the gap had a length that was not a multiple of 3, and its length was <30 bp long (to remove probable introns from consideration), the gap was determined to be a frameshift and its location saved to a BED file.

### Nonsense mutations

We also analysed the number of nonsense mutations that would cause early stop codons in both the Ensembl and CAT gene annotation sets. We identified nonsense mutations by iterating through each codon in the coding sequence of the predicted transcripts. If there was an early stop codon

before the canonical stop codon at the end of the transcript, we saved the location in a BED file.

### Validation of mutations using Illumina

For both sets of mutations, we then lifted over the coordinates of the mutations to be on the GRCh38 reference so that we could use existing variant call sets on GRCh38. We used halLiftover to lift over each set of coordinates, using the GRCh38-based HAL file from the MC alignment. Then we used bedtools intersect to intersect with the variant call file for each of the assemblies.

The following sample commands were used:

halLiftover GRCh38-f1g-90-mc-aug11.hal <GENOME_NAME> <MUTATION_BED_FILE> GRCh38 <LIFTED_OVER_BED_FILE>
bedtools intersect -wo -a <LIFTED_OVER_BED_FILE> -b <SAMPLE_MERGED_VCF> > <OVERLAP_OUTPUT_TXT_FILE>

The VCF files used in this intersection were downloaded from the 1KG (https://urldefense.com/v3/__https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_raw_GT_with_annot/20201028_CCDG_14151_B01_GRM_WGS_2020-08-05_chr$i.recalibrated_variants.vcf.gz__;!!NLFGqXoFfo8MMQ!r6nD4EtteJ7k2BauOrREfgIrlxEI2Upx35sNHiqyI8Did-a6UUUzzxGVQkwYkb-bE_rlHQN2Jw2cBdlw7te_-Q$).

Where $i was replaced with each chromosome number. From there, each chromosome VCF was split so that each sample was in its own file using bcftools view. The chromosome files for each sample were combined into one VCF using bcftools concat.

## Gene duplication analysis

Duplicated genes were detected as multi-mapped coding sequences using Liftoff[88] supplemented by a complementary approach (gb-map) with multi-

mapped gene bodies. The combined set was formed by including all Liftoff gene duplications and duplicated genes detected by gb-map.

**Liftoff**

We ran Liftoff (v.1.6.3) to annotate extra gene copies in each of the assemblies. Liftoff was run with the flag -sc = 0.90 to find additional copies of genes, with an identity threshold of at least 90%. An example command is below:

```
liftoff -p 10 -sc 0.90 -copies -db <GENCODE_V38_DATABASE> -u
<UNMAPPED_FILE> -o <OUTPUT_GFF3> -polish
<GENOME_FASTA> <GRCh38_FASTA>
```

The additional copies of the genes were identified as such in the output gff3 with the field extra_copy_number (equal to anything other than 0). For this analysis, we also only considered genes that were multi-exon, protein-coding genes. The additional gene copies were further filtered to remove any genes outside the 'reliable' haploid regions as determined by the Flagger pipeline.

**gb-map**

The gene-body mapping pipeline identifies duplicated genes by first aligning transcripts of protein-coding and pseudogenes (GENCODE v.38) to each assembly and then multi-mapping the genomic sequences of each corresponding gene. Alignments of at least 90% identity and 90% of the length of the original duplication were considered candidate duplicated genes. Candidates were removed if they overlapped previously mapped transcripts from other genes, low-quality duplications and genes identified through CAT and Liftoff analysis.

**Gene family analysis**

To account for gene duplications in high-identity gene families, gene families were identified on the basis of sequence alignments from gb-map.

Genes that mapped reciprocally with 90% identity and 90% length were considered a gene family. A single gene was selected as the representative gene for the family, and any gene duplication in the family was counted towards that gene.

## Pangenome graph construction

### Minigraph

Minigraph can rapidly perform assembly-to-graph mappings using a generalization of the minimap2 algorithm[34]. New SVs of at least 50 bp detected in the mapping can then be added to the graph. To construct a pangenome graph, one chosen reference assembly, GRCh38 in this case, was used as a starting graph, and the mapping and SV addition steps were repeated for each additional assembly, greedily. This iterative approach is analogous to partial order alignment (POA)[89]. Graphs constructed in this way describe the structural variation within the samples and provide a coordinate system across the reference and all insertions. Minigraph does not produce self-alignments. That is, it will never align a portion of the reference assembly onto another portion of the reference assembly. In this way all reference positions have a unique location within the created pangenome. Minigraph (v.0.14) was used with -xggs options. The input order was GRCh38, CHM13 then the remainder in lexicographic order by sample name.

### MC

Graphs constructed by Minigraph only contain structural variation (≥50 bp) by default. The aim of the MC pipeline is to refine the output of Minigraph to include smaller variants, down to the SNP level. Doing so allows the graph to comprehensively represent most variation, as well as to embed the input haplotypes within it as paths, which is important for some applications[49]. To remove noisy alignments from the MC pangenome, long (≥100 kb) non-reference sequences identified as being satellite, unassignable to a reference chromosome or which appear unaligned to the remainder of the assemblies were removed from the graph. This resulted in

a pangenome with significantly reduced complexity that nevertheless maintained all sequences of the starting reference assembly and the large majority of those in the additional haplotypes. The MC pipeline is composed of the five steps described below and in more detail in ref. [35]. The script and commands to reproduce this process can be found at GitHub (https://github.com/ComparativeGenomicsToolkit/cactus/blob/81903cb82ae80da342515109cdee5a85b2fde625/doc/pangenome.md#hprc-version-10-graphs). A newer, simpler version of the pipeline that no longer requires satellite masking can be found at GitHub (https://github.com/ComparativeGenomicsToolkit/cactus/blob/5fed950471f04e9892bb90531e8f63be911857e1/doc/pangenome.md#hprc-graph).

Paths from the reference, GRCh38, are acyclic in the MC graph. Paths from any other haplotypes can contain cycles (as a result of different query segments mapping to the same target), but they are relatively rare.

1. (1)

   Satellite masking: Minigraph is unable to map through highly repetitive sequences such as centromeres and telomeres and, as these regions are also enriched for misassemblies (see the section 'Assembly assessment' of Results), we decided to exclude them from the MC graphs used in this work. dna-brnn is a tool that uses a recurrent neural network to quickly identify alpha satellites as well as human satellites 1 and 2 (ref. [90]). We ran it with its default parameters on all input sequences and cut out any identified regions ≥100 kb, except on the reference. The three satellite families that dna-brnn detects account for the majority of satellite sequences, but not all. As such, gaps ≥100 kb in Minigraph mappings were also removed. They were detected by mapping each assembly, after having removed the dna-brnn regions, to the Minigraph (using the procedure described below). Overall, an average of 188.6 Mb of sequence from each (non-reference) assembly was excluded from the graph.

2. (2)

   Assembly-to-graph mapping: Minigraph generalizes the fast seeding and chaining algorithms of minimap2, but it does not currently

produce exact alignments in cigar strings or otherwise. For this work, an option, --write-mz, was added to report chains of minimizers, which in this case are 15 bp exact matches, and all assemblies were mapped to the Minigraph graph using it. The resulting minimizers were then converted into PAF files with cigars representing exact pairwise alignments between the query contigs and Minigraph node sequences, and all mappings with MAPQ < 5 were excluded, as were overlapping query regions >10 kb.

3. (3)

   Chromosome decomposition: the Minigraph graphs do not contain inter-chromosomal rearrangements, but the mappings performed in the previous steps can imply them. That is, a contig can partially map to multiple chromosomes. In most cases, these mappings involve similarity across different acrocentric short arms. To avoid introducing misleading interchromosomal events, and because it is necessary to run the subsequent steps individually on chromosomes owing to memory requirements, the mappings were divided by reference chromosome. This was done by splitting the Minigraph into connected components and using the RN tags to determine their corresponding chromosome names. The PAF mappings were used to determine the coverage of each query contig with each chromosome component. This coverage was used to assign each query contig to a single chromosome by choosing the chromosome with the highest coverage. Contigs with insufficient coverage to any chromosome (<90% for contigs with lengths in [1,10 kb); <80% for [10 kb,100 kb), <75% for [100 kb,1 Mb) and <70% for ≥mb.) were considered ambiguous and not included in the graph. In the GRCh38-based graph, all unplaced and random contigs were grouped together into the same component.

4. (4)

   Cactus base alignment: Cactus is a tool that uses a graph-based approach to combine sets of pairwise alignments obtained from lastz into a multiple genome alignment[37]. When aligning different species, it uses a phylogenetic tree to progressively decompose the alignment into a subproblem for each ancestral node in the tree. We adapted it to also

accept chromosome-scale sequence-to-Minigraph mappings as produced above, and improved its runtime on alignments of many sequences by replacing its base aligner with abPOA[91]. The core algorithm described in ref. [37] remains unchanged, whereby the pairwise alignments were used to induce a sequence graph, then filtered using the Cactus alignment filtering algorithm, and components of unaligned sequence were then processed by the base alignment and refinement algorithm. The resulting graph was used to infer an ancestral sequence (not explicitly used in this work) and then exported to a hierarchical alignment (HAL) file[92]. We implemented a converter, hal2vg[93] that converts the HAL alignment into a sequence graph in VG format.

5. (5)

   Post-processing and whole-genome indexing: the following post-processing steps were performed on each chromosome graph. First, unaligned sequences >10 kb in length, including sequences not aligned to Minigraph, were removed to filter out any under-alignment artefacts that might later be mistaken for insertions. Next, GFAffix[94] was used to normalize the graphs by merging together redundant node prefixes and suffixes. Nodes were flipped as necessary to ensure that reference paths always visit their forward orientations. The chromosomes were combined into a whole-genome graph, indexed and exported to VCF, all using vg. Patched versions of both the GRCh38-based and CHM13-based graphs were created when it was discovered that short contigs split-mapping to distant locations had induced large deletions. The deletions were removed using vg clip -D 10000000 (and the pipeline has since been corrected to no longer produce them). Allele-filtered graphs, used for short-read mapping, were produced (from the patched graphs) by removing all nodes traversed by fewer than 9 haplotype paths (minimum AF = 10%) using vg clip -d 9 -m 10000. The chromosome HAL files were also combined into a whole-genome HAL file using halMergeChroms, and clipped sequences added back (to facilitate running CAT) using halUnclip.

#### PGGB

The PGGB uses a symmetric, all-to-all comparison of genomes to generate and refine a pangenome. We applied it to build a pangenome graph from all genome assemblies and references (both GRCh38 and CHM13). The resulting PGGB graph represents all alignment relationships between input genomes in a single graph. The PGGB graph is a lossless model of the input assemblies that represents all equivalently. This arrangement enables all of our pangenome assemblies to be used as reference systems, a property that we used to explore the scope of pangenome variation in a total way. Owing to ambiguous placement of variation in all-to-all pairwise alignments, many SV hotspots, including the centromeres, are transitively collapsed into loops through a subgraph representing a single repeat copy, a feature that tends to reduce the size of variants found in repetitive sequences. In contrast to MC, PGGB does not filter rapidly evolving satellite sequences or the regions that do not reliably align. This increases its size and complexity relative to the MC graph and adds a significant amount of singleton sequences relative to the Minigraph and MC graphs. However, this property enables annotations and coordinates of all assemblies in the pangenome to be related to the graph structure and utilized in subsequent downstream analyses. We applied the PGGB model to investigate the full pangenome and integrate annotations established de novo on the diverse assemblies into a single model for analyses of pangenome diversity and of complex structurally variable loci (MHC and 8p inversion).

PGGB generates a pangenome graph in three phases. (1) Alignment: in the first phase, the wfmash aligner[95] is used to generate all-vs-all alignments of input sequences. This method, wfmash, applies the mapping algorithm of MashMap2 to find homologies at a specified length and per cent identity. It then derives base-level alignments using a high-order version of the WFA algorithm (wflign), which first aligns sequences in segments of 256 bp, then patches up the base-level alignment with local application of WFA. wfmash was designed and developed specifically for the problem of building all-to-all alignments for large pangenomes. (2) Graph induction: the input FASTA sequences and PAF-format alignments produced by wfmash are converted to a graph (in GFA format) using seqwish[96]. This losslessly transforms the input alignments and sequences into a graph. (3) Graph normalization: we applied a normalization algorithm—smoothxg[97]—to simplify complex motifs that occur in STRs and other repetitive sequences, as well as to

mitigate underalignment. The graph is first sorted using a path-guided stochastic gradient descent method[98] that organizes the graph in one-dimension to optimize path distances and graph distances. This sort provides a way to partition the graph into smaller pieces over which we applied a multiple sequence alignment algorithm (abPOA)[91]. These pieces were laced back into a final graph. We iterated this process twice using different target POA lengths to remove boundary effects caused at the borders of the MSA problems. Finally, we applied GFAffix[94] to remove redundant furcations from the topology of the graph.

To build the HPRC PGGB graph, we used both the CHM13 and GRCh38 references as a target and mapped all contigs against these with wfmash, requiring a full length mapping at 90% total identity, collecting all contigs that mapped to a given chromosome. Contigs that did not map under this arrangement were then partitioned using a split-mapping approach, requiring 90% identity over 50 kb to seed the mappings, and putting the contig into the chromosome bin for which it had the best split mapping. We thus initially partition the data into 25 chromosome sets: one for each autosome, one for each sex chromosome, and finally the mitochondria.

We then applied PGGB (v.0.2.0+531f85f) to each partition to build a chromosome-specific graph. Run in parallel over 6 PowerEdge R6515 AMD EPYC 7402P 24-core nodes with 384 GB of RAM, this process requires 22.49 system days, or around 3.7 days wallclock. To develop a robust process to build the HPRC graph, the PGGB team iterated the build 88 times. The final chromosome graphs were compacted into a single ID space using vg ids -j, then for each reference (GRCh38 and CHM13) a combined VCF file was generated from the graph with vg deconstruct (v.1.36.0/commit 375cad7).

A handful of key parameters defined the shape of the resulting graph. First, in wfmash, we required >100 kb mappings at 98% identity. We mapped each HPRC assembly contig and reference chromosome (both GRCh38 and CHM13) to all the other 89 input haplotypes. To reduce complexity, and false-positive SNPs resulting from misaligned regions, we applied a minimum match length filter (in seqwish) of 311 bp. This meant that the graph that we induced was relatively ‘underaligned’ locally, and only

through normalization in smoothxg did we compress the bubble structures that are produced. For smoothxg, our first iteration attempts to generate 13,033 bp-long POA problems, whereas the second is 13,177 bp. These lengths provided a balanced trade-off between run time and variant detection accuracy.

In addition to a graph (in GFA), PGGB generates visualizations of the graph in one and two dimensions, which show both the topology (two dimensional) and path-to-graph relationship (one dimensional). A code-level description of the build process is provided at GitHub (https://github.com/pangenome/HPRCyear1v2genbank).

## Pangenome graph assessment

### Annotating variant sites in pangenome graphs

Variant sites in Minigraph and in MC and PGGB graphs were discovered using gfatools bubble (v.0.5)[75] and vg deconstruct[99], respectively. Large (>10 Mb) spurious deletions in MC and PGGB graphs were removed using vcfbub (v.0.1.0)[100] with options -l 0 -r 10000000. Next, variant sites were classified into small variant (<50 bp) and SV (≥50 bp) sites. The SV sites were then annotated as described in the methods section of article that describes Minigraph[34]. In brief, the longest allele sequence of each SV site was extracted and stored in the FASTA format. The interspersed repeats, low-complexity regions, exact tandem repeats, centromeric satellites and gaps in the longest allele sequences were then identified using RepeatMasker (v.4.1.2-p1) with the NCBI/RMBLAST (v.2.10.0) search engine and Dfam (v.3.3) database, SDUST (v.0.1)[101], ETRF (commit fc059d5)[102], dna-brnn (v.0.1)[90] and seqtk gap (v.1.3)[103], respectively. SDs were identified if the total node length in a site was ≥1,000 bp and ≥20% of bases of these nodes were annotated as SD in the reference or in individual assembly ('SD annotation' subsection). To find hits to the GRCh38 reference genome, minimap2 (v.2.24) with options -cxasm20 -r2k --cs was used to align the longest allele sequences to the reference genome. Based on the identified features, SV sites were classified into various repeat classes using mgutils.js anno (https://github.com/lh3/minigraph/blob/master/misc/mgutils.js) with minor modifications to enable it to work with the files derived from the MC and PGGB graphs.

### Pangenome size and growth

We use the heaps tool of the odgi pangenome analysis toolkit[98] to estimate how the euchromatic autosomal pangenome grows with each additional genome assembly added. Here we approximated euchromatic regions by non-satellite DNA, which was identified by dna-brnn in the construction of the MC graph (see the 'MC' subsection). Although the MC non-reference

haplotypes of the MC graph do not contain satellite DNA, the PGGB graph does. Consequently, we subset the PGGB graph to segments contained in the MC graph. We additionally excluded reference haplotypes (GRCh38 and CHM13) from the analysis. We then sampled permutations of the 88 non-reference (neither GRCh38 nor T2T-CHM13) haplotypes. In each permutation, we calculated the size of the pangenome after adding the first 1, 2, …, *N* haplotypes in both graphs. This produced a collection of saturation curves from which we derived a median saturation curve onto which we fitted a power law function known as Heaps' Law. The exponent of this function is generally understood to represent the degree of openness —or diversity—of a pangenome[39]. Summing up, we called odgi heaps -i <graph.gfa> -S -n200 to generate pangenome saturation curves for 200 permutations. Next to calculating a non-permuted cumulative base count, we also counted the number of common (≥5% of all non-reference haplotypes) and core (≥95% of all non-reference haplotypes) bases in the pangenome graphs. To this end, we used a tool called panacus[104] and supplied a list of the samples in which they are grouped according to their assigned superpopulation (pangenome-growth -m -t bp <graph.gfa> <sample order>). We repeated the count, this time including only segments of depth ≥2, that is, contained at least twice in any haplotype sequence.

**Decomposing pangenome graphs based on allele traversals**

Pangenome graphs were decomposed topologically into a set of nested subgraphs, termed snarls, that each correspond to one or a collection of genetic variants. These snarls were then converted to VCF format using vg deconstruct[99]. Large (>100 kb) deletions in MC and PGGB graphs were removed using vcfbub (v.0.1.0)[100] with options -r 100000. To ease the comparison of variants with other call sets for each individual, the multi-sample VCF files were converted to per-sample VCF files using bcftools view -a -I -s <sample name>, and the multiallelic sites were split into biallelic records using bcftools norm -m -any. Owing to the limitations of snarl decomposition, snarls may contain multiple variants that cannot be further decomposed into nested snarls using vg deconstruct. If snarls of this kind are compared with truth calls, the evaluation will not be accurate. We solved this problem by comparing reference and alternate allele traversals

for each snarl to infer the minimalist representation of variants (Supplementary Fig. 44).

### Annotating small variants in pangenome graphs with AFs from gnomAD

Variant sites in MC and PGGB graphs were discovered using vg deconstruct[99]. The resulting VCF files were then decomposed on the basis of allele traversals (‘Decomposing pangenome graphs based on allele traversals’ subsection). The multi-nucleotide polymorphisms and complex indels were further decomposed into SNPs and simple indels using vcfdecompose --break-mnps --break-indels from RTG tools (v.3.12.1)[83], so that they could be annotated with gnomAD later. For comparison, variants called from PacBio HiFi reads using DeepVariant and from haplotype-resolved assemblies using Dipcall were also used. For each discovery method, small variants (<50 bp) were extracted and normalized using bcftools norm -c s -f <reference sequence in FASTA format> -m -any. Next, all per-sample VCF files were combined into one VCF file using bcftools concat -a -D after dropping individual genotype information using bcftools view -G. To annotate small variants with AFs from gnomAD[105], the gnomAD (v.3.1.2) per-chromosome VCF files were downloaded and concatenated into one VCF file using bcftools concat. The VCF file was then compressed into a file in the gnotate format using make-gnotate from slivar (v.0.2.7)[106] with options --field AC:gnomad_ac --field AN:gnomad_an --field AF:gnomad_af --field nhomalt:gnomad_nhomalt. The small variants were annotated with gnomAD using slivar expr --gnotate <gnotate file>.

## Variant benchmarking

### Calling variants from PacBio HiFi reads

The PacBio HiFi reads were aligned to the GRCh38 human reference genome with no alternatives using Winnowmap2 (v.2.03)[107] with -x map-pb -a -Y -L --eqx --cs. The MD tags required by Sniffles were calculated using

samtools calmd. The resulting BAM files were sorted and indexed using SAMtools.

For small variants, the two-pass mode of DeepVariant (v.1.1.0)[107] with WhatsHap (v.1.1)[108] was used to call SNPs and indels from the PacBio HiFi read alignments. The resulting VCF files were used as truth sets for small variant benchmarking.

Three discovery methods were used to call SVs from the PacBio HiFi read alignments. For PBSV (v.2.6.2)[41], SV signatures were identified using pbsv discover with --tandem-repeats <GRCh38 TRF BED file> to improve the calling performance in repetitive regions. SVs were then detected using pbsv call with --ccs --preserve-non-acgt -t DEL,INS,INV,DUP,BND -m 40 from the signatures. For SVIM (v.2.0.0)[44], SVs were called using svim alignment with --read_names --zmws --interspersed_duplications_as_insertions --cluster_max_distance 0.5 --minimum_depth 4 --min_sv_size 40. In contrast to PBSV and Sniffles, SVIM outputs all calls no matter their quality. To determine the threshold used for filtering low-quality calls, a precision–recall curve was generated across various quality scores by comparing with the GIAB (v.0.6) Tier 1 SV benchmark set for HG002 (Supplementary Fig. 45). Consequently, SVIM calls with a quality score lower than ten were excluded. For Sniffles (v.1.0.12b)[42], SVs were discovered with -s 4 -l 40 -n -1 --cluster --ccs_reads. Unlike PBSV and SVIM, Sniffles does not generate consensus sequences of insertions from aggregating multiple supporting reads. Therefore, Iris (v.1.0.4)[43] was used to refine the breakpoints and insertion sequences with --hifi --also_deletions --rerunracon --keep_long_variants. All resulting VCF files were sorted and indexed using BCFtools.

**Calling SVs from haplotype-resolved assemblies**

Three discovery methods were used to call SVs from the haplotype-resolved assemblies generated using Trio-Hifiasm.

For SVIM-asm (v.1.0.2)[45], assemblies were aligned to the GRCh38 human reference genome with no alternatives using minimap2 (v.2.21)[86] with -x

asm5 -a --eqx --cs and then sorted and indexed using SAMtools. SVs were called using svim-asm diploid with --query_names --interspersed_duplications_as_insertions --min_sv_size 40. The resulting VCF files were sorted and indexed using BCFtools.

For PAV (v.0.9.1)[5], assemblies were aligned to the GRCh38 human reference genome with no alternatives using minimap2 (v.2.21)[86] with options -x asm20 -m 10000 -z 10000,50 -r 50000 --end-bonus=100 --secondary=no -a --eqx -Y -O 5,56 -E 4,1 -B 5. These alignments are then trimmed to reduce the redundancy of records and to increase the contiguity of alignments. SVs, indels and SNPs were called by using cigar string parsing of the trimmed alignments. Inversion calling in PAV uses a new *k*-mer density assessment to resolve inner and outer breakpoints of flanking repeats, which does not rely on alignment breaks to identify inversion sites. This is designed to overcome limitations in alignment methodologies and to expand inversion calls, which result in duplications and deletions of sequence on the boundaries.

The Hall-lab pipeline is as documented in the WDL workflow (https://github.com/hall-lab/competitive-alignment/blob/master/call_assembly_variants.wdl) (commit 830260a). In brief, the maternal and paternal assemblies were aligned to the GRCh38 human reference genome using minimap2 (v.2.1)[86] with options -ax asm5 -L --cs. Large indels (>50 bp) were detected using the 'call_small_variants' task, based on paftools (v.2.17-r949-dirty). For large SV, breakpoints were mapped based on split alignments of assembly contigs to the reference genome and classified as SVs using a series of custom python scripts in the 'call_sv' task. The breakpoint-mapped SVs were then filtered on the basis of the coverage of the reference genome by the assembly contigs (calculated using bedtools genomecov (v.2.28.0)). For each haplotype assembly, a BED file of 'excluded regions' was defined comprising genomic regions covered by more than one distinct contig or with more than 3× coverage by a single contig. Breakpoint-mapped SVs where either breakpoint or >50% of the outer span intersected an excluded region were filtered.

## Merging SV call sets

To integrate per-sample VCF files generated by three HiFi-based and three assembly-based SV callers, svtools[109] was used. For each individual, VCF files from the six callers were jointly sorted and then merged using svtools lsort and lmerge, first using a strict criterion (svtools lmerge -f 20), followed by a more lenient second merge (svtools lmerge -f 100 -w carrier_wt). The autosomal SV calls supported by at least two callers were included in the consensus SV call set for comparison.

**Defining confident regions for variant benchmarking**

For SVs, confident regions were generated using Dipcall. Although useful for small variants, current benchmarking tools such as hap.py/vcfeval cannot properly compare different representations of small variants in and around SVs. Therefore, for each sample, the confident regions from Dipcall were further processed as follows:

1. (1)

   Exclude any SD, self-chain, tandem repeat longer than 10 kb or satellite DNA if there are any breaks in the Dipcall BED file in the repeat region +15 kb flanking sequence on each side. The rationale is that breaks in the Dipcall BED file are generally caused by missing sequence or errors in the assembly or reference or by large SVs or CNVs for which we do not have tools to benchmark small variants in these regions.

2. (2)

   Exclude 15 kb around all breaks in the Dipcall BED file for the same reason as noted above.

3. (3)

   Exclude 15 kb around all gaps in GRCh38 because alignments are unreliable.

4. (4)

Exclude variants >49 bp in the Dipcall VCF file and any tandem repeats overlapping these SVs +50 bp on each side.

### Benchmarking variants

Variant sites in MC and PGGB graphs were discovered using vg deconstruct[99]. Variant sites with alleles larger than 100 kb in MC and PGGB graphs were then removed using vcfbub (v.0.1.0)[100] with options -l 0 -a 100000. The resulting VCF files were further processed using vcfwave from vcflib[110] with option -I 1000. In brief, vcfwave realigned alternate alleles against the reference allele for each variant site using the bidirectional wavefront alignment (BiWFA) algorithm[111] to decompose complex alleles into primitive ones. The multi-sample VCF files were then converted to per-sample VCF files using bcftools view -a -I -s <sample name> and the multiallelic sites were splitted into biallelic records using bcftools norm -m -any. Next the autosomal small variants (<50 bp) from a given pangenome graph (query set) were compared with the HiFi-DeepVariant call set (truth set) using vcfeval from RTG tools (v.3.12.1)[83] with options -m annotate --all-records --ref-overlap --no-roc. Note that the multi-nucleotide polymorphisms and complex indels were reduced to SNPs and simple indels using vcfdecompose --break-mnps --break-indels from RTG tools (v.3.12.1)[83] The comparison was performed independently for each individual. Recall and precision were calculated within the refined Dipcall confident regions (‘Defining confident regions for variant benchmarking’ subsection) and then stratified using the GIAB (v.3.0) genomic context. To evaluate the SV (≥50 bp) calling performance, the autosomal SVs from a given pangenome graph (query set) were compared to the consensus SV call set (truth set) for each individual using truvari bench (v.3.2.0)[112] with options --multimatch -r 1000 -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15 --sizemax 100000 --includebed <Dipcall confident regions>. Recall and precision were then stratified using the GIAB (v.3.0) genomic context and by variant length.

## Alignment of long reads to pangenomes

### PacBio HiFi reads

PacBio HiFi reads from 44 HPRC samples (excluding the held out samples) were aligned to the MC graph using GraphAligner (v.1.0.13)[113] with option -x vg and stored in the GAF format[34]. For each read that aligned to multiple places in the graph, the alignment with the highest score was retained. To remove low-quality alignments, a read with <80% of read length aligned to the graph was discarded. After filtering the read-to-graph alignments, the read depth of each edge was calculated using vg pack (v.1.33.0)[114] with options -Q -1 -D. Note that the resulting GAF files did not contain a mapping quality (encoded as 255 for missing) for each alignment, therefore the option -Q -1 was given to vg pack to ensure that these alignments were used during read-depth calculation. Next, the edges of each sample were classified into either on-target or off-target depending on whether they were on the sample paths (encoded as W-lines in MC GFA files) or not.

### Oxford Nanopore reads

ONT reads obtained from 29 HPRC samples (samples labelled HPRC in Supplementary Table 1) were aligned against the MC graph. The alignments were produced using GraphAligner (v.1.0.13) with parameter settings -x vg --multimap-score-fraction 1 --multiseed-DP 1. The number of reads in these datasets range between 1 million and 5.4 million and have an average read length of 28.4 kb. On average, 99.68% of the reads received hits from one or more locations in the graph. For each read, we determined its best hit based on alignment score and discarded all its lower-scoring alignments in subsequent analyses. The alignment identities of these best hits peaked above 95%, with an average ratio of alignment-length-to-read-length (ALRL) of 0.880 (s.d. = 0.302) and average MAPQ value of 59.35. The alignment set was further quality-pruned by discarding alignments that either had an ALRL lower than 0.8 or a MAPQ value lower than 50. The surviving alignments had an overall average ALRL of 0.968 (s.d. = 0.047) and effectuate an overall genome coverage between 10.5-fold and 43-fold across the 29 samples (Supplementary Fig. 46).

## Annotating genes within pangenomes

We ran CAT[46] to annotate each of the genomes within a pangenome graph. CAT projects a reference annotation, in this case GENCODE (v.38), to each of the haplotypes using the underlying alignments within the graph. CAT (commit eb2fc87) was run on both of the GRCh38-based and the CHM13-based MC graphs . For each graph, the autosomes were first run all together, and then the sex chromosomes were run on the appropriate haplotypes. The parameters used were default parameters, except as shown below. An example CAT command run is:

luigi --module cat RunCat --hal=CHM13-f1g-90-mc-aug11.hal --ref-genome=GRCh38 --workers=8 --config=cat-hprc.gencode38.autosomes.config --work-dir work-hprc-gencode38-chm13 --out-dir out-hprc-gencode38-chm13 --local-scheduler --assembly-hub --maxCores 8 --binary-mode local > cat.hprc.gencode38.autosomes.chm13.log

Comparisons were made between the resulting CAT annotations and those from the Ensembl pipeline by looking at the parent GENCODE identifiers for each gene and transcript in the sets. Numbers of shared and unique identifiers between the sets were tabulated. Because the two annotation sets used slightly different versions of GENCODE (v.38), only those identifiers attempted to be mapped by both pipelines were considered. Additionally, features were considered to be at the same locus if their genomic intervals overlapped.

## Identifying medically relevant sites

SV sites in Minigraph were discovered using gfatools bubble. To obtain the number of observed alleles per site, per-sample alleles were called using minigraph -cxasm --call (Supplementary Table 21). SV sites with alleles larger than 10 kb and at least five observed alleles were selected as complex SV sites. The complex SV sites were further filtered on the basis of whether they overlapped with medically relevant protein-coding genes[47] using bedtools intersect. To understand whether the medically relevant complex SV sites are known in previous studies, the coverage of SVs from the 1KG call sets[10,19] was computed using bedtools coverage -F 0.1 (Supplementary

Table 16). All complex SVs were examined using Bandage[115] and visually compared to previous short-read SV call sets[10,19,68,116] using IGV.

## Analysis of five complex loci

### Visualization of graph structures of five loci

We extracted subgraphs and paths for five loci in the MC and PGGB graphs using gfabase (v.0.6.0)[117] and odgi (v.0.6.2)[98] with the following example commands:

gfabase sub GRCh38-f1g-90-mc-aug11.gfab GRCh38.chr1:25240000-25460000 --range --connected --view --cutpoints 1 --guess-ranges -o RH_locus.walk.gfa odgi extract -i chr1.pan.smooth.og -o chr1.pan.RH_locus.og -b chr1.RH_locus.bed -E -P

We then visualized the graph structures of the subgraphs using Bandage (v.0.8.1)[115].

### Alignment of genes to graphs

We aligned Ensembl (release 106)[77] GRCh38 version gene sequences to the MC graph and PGGB graph using GraphAligner (v.1.0.13)[113] with parameter settings -x vg --try-all-seeds --multimap-score-fraction 0.1 to identify the gene positions within the graphs. To show locations of genes on Bandage plots, we applied colour gradients from green to blue to the nodes of each gene. Lines alongside the Bandage plots showing approximate gene positions, exons and transcription start sites based on Ensembl Canonical transcripts were drawn by hand.

### Structural haplotypes identification

Sequences of each assembly are represented by paths in a GFA file. We identified SVs in each assembly by tracing these paths through different ‘big’ bubbles (>5,000 bp) in either the MC graph or PGGB graph within those gene regions. We selected the 5,000 bp bubble size based on manual

inspection of Bandage plots. An example command to identify big bubbles at a RH locus is as follows:

```
bcftools filter hprc-v1.0-mc-grch38.vcf.gz -r chr1:25240000-25460000 |
grep LV=0 | awk '{OFS="\t"; print $1,$2,$3,$4,$5}' | tr "," "\t" | awk
'{OFS="\t"; for (i=1;i<=NF;i++) {len=length($i); if (len>5000) {print
$1,$2,$3; next}}}'
```

To identify gene conversion events (as gene conversions are not shown as bubbles in the graphs), we identified nodes that were different between a gene and its homologous gene (for example, *RHD* and *RHCE*) based on the GraphAligner alignments described above. We refer to these as paralogous sequence variants. A gene conversion event was detected if a path of a gene goes through more than four paralogous sequence variants of its homologous gene in a row.

### Visualization of linear gene structures

We counted the number of assemblies for each structural haplotype and computed their frequency. We visualized linear haplotype structures (for example, in Fig. 5c) using gggenes (v.0.4.1)[118] based on the structural haplotypes determined for each assembly from the pangenome graphs. The length of intervals between genes is fixed (except for *TMEM50A* and *RHCE*, because those two genes are immediately next to each other). Lengths of genes are shown as proportional to gene lengths in GRCh38.

## Pangenome point genotyping

### Alignment of reads to the pangenome

The short reads were first split into chunks to parallelize the read mapping to the 'allele-filtered graph' pangenome, as defined in the 'MC' subsection. This pangenome is included within the dataset accompanying this paper and can be identified as 'clip.d9.m1000.D10M.m1000'. Mapping was performed with Giraffe[49] from vg release v.1.37.0. For trio-based runs, the trio-sample sets of short reads were mapped to the pangenome using Giraffe

from vg release v.1.38.0. Note that the core vg algorithms for Giraffe mapping and surjection (conversion from graph space to linear space) are the same in both vg v.1.37.0 and v.1.38.0. The output alignments, surjected to GRCh38 in BAM format as explained below, are available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/PANGENOME_2022/DeepTrio/samples in the bam directory of each sample's directory, and are organized by aligner.

## Surjection to GRCh38 and indel realignment

To perform variant calling, GAM alignments were surjected onto the chromosomal paths from GRCh38 (chromosomes 1–22, X and Y) using vg surject and the --prune-low-cplx option to prune short and low-complexity anchors during realignment. The BAM files were sorted and split by chromosome using SAMtools (v.1.3.1)[119]. The reads were realigned, first using bamleftalign from FreeBayes (v.1.2.0)[120], and then with ABRA (v.2.23)[121] on target regions that were identified using RealignerTargetCreator from GATK (v.3.8.1)[122] and expanded by 160 nucleotides with bedtools slop (v.2.21.0)[123].

## Model training

To perform variant calling with DeepVariant and DeepTrio, we trained machine-learning models specific to our graph reference and Giraffe alignment pipeline based on our alignments. For all models, chromosome 20 was entirely held out from all input samples to provide a control.

Training was performed on Google's internal cluster, using unreleased Google tensor processing unit (TPU) accelerators, from a cold start (that is, without using a pre-trained model as input). We believe that nothing about the way in which we executed the training is essential to the results obtained. Cold start training is estimated to be feasible outside the Google environment; therefore the claims we present here are falsifiable, but it is not expected to be cost-effective. Researchers looking to independently replicate our training should consider doing warm start training from a base model trained on other data, using commercially available graphics

processing unit (GPU) accelerators. An example procedure can be found in the DeepVariant training tutorial at GitHub (https://github.com/google/deepvariant/blob/r1.3/docs/deepvariant-training-case-study.md). We predict that this more accessible method would produce equivalent results.

For both DeepVariant and DeepTrio, the true variant calls being trained against came from the GIAB benchmark (v.4.2.1).

For DeepVariant, we trained on the HG002, HG004, HG005, HG006 and HG007 samples, with HG003 held out. The trained DeepVariant model is available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/PANGENOME_2022/DeepVariant/models/DEEPVARIANT_MC_Y1.

For DeepTrio, we trained two sets of models: one on HG002, HG003, HG004, HG005, HG006 and HG007, with HG001 held out; and one on HG001, HG005, HG006 and HG007, with the HG002, HG003 and HG004 trio held out. Each DeepTrio model set included parental and child models. The two trained child deeptrio models are available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/PANGENOME_2022/DeepTrio/models/deeptrio/child and https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/PANGENOME_2022/DeepTrio/models/deeptrio-no-HG002-HG003-HG004/child, respectively. The two trained parental DeepTrio models are available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/PANGENOME_2022/DeepTrio/models/deeptrio/parent and https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/PANGENOME_2022/DeepTrio/models/deeptrio-no-HG002-HG003-HG004/parent, respectively.

**Variant calling with DeepVariant**

DeepVariant (v.1.3) was evaluated on HG003, using the model we trained with HG003 held out (see 'Model training'). We used the --

keep_legacy_allele_counter_behavior flag (introduced to support this analysis) and a minimum mapping quality of 1 in the make_examples step, before calling the variants with call_variants. Both VCFs and gVCFs were produced. The WDL workflow used for single sample mapping and variant calling was deposited into Zenodo (https://doi.org/10.5281/zenodo.6655968).

### Variant calling on GRCh38 with BWA-MEM and DeepVariant

Small variants were also called using a more traditional pipeline. We aligned reads with BWA-MEM[50] to GRCh38 with decoys but no ALTs. DeepVariant then called small variants from the aligned reads. The same version and parameters were used for DeepVariant. Only the model was changed to the default DeepVariant model.

### Variant calling with DeepTrio

Small variants were also called using DeepTrio (v.1.3). For HG001, we used the DeepTrio models we trained with HG001 held out (see 'Model training'). For the HG002, HG003 and HG004 trio and HG005, HG006 and HG007 trio, we used the models trained with the HG002, HG003 and HG004 trio held out; the HG005, HG006 and HG007 trio (except for chromosome 20) was still included in the training set. We used the --keep_legacy_allele_counter_behavior and a minimum mapping quality of 1 in the make_examples step before calling the variants with call_variants. Both VCFs and gVCFs were produced and are available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/PANGENOME_2022/DeepTrio/samples in the vcf directory of each sample's directory, and are organized by mapping and calling condition. The WDL workflow used for trio-based mapping and variant calling was deposited into Zenodo (https://doi.org/10.5281/zenodo.6655962).

### Variant calling on GRCh38 with BWA-MEM, Dragen Graph and DeepTrio

For DeepTrio, small variants were also called using a more traditional pipeline and a graph-based implementation of Illumina's Dragen platform (v.3.7.5). The conditions evaluated were each a combination of a mapper and a reference. The Giraffe-HPRC condition used Giraffe (v.1.38.0)[49] to align reads to the HPRC reference. The BWA-MEM condition used BWA-MEM (v.0.7.17-r1188)[50] to align reads to the hs38d1 human reference genome with decoys but no ALTs. The Dragen-DeepTrioCall condition used Illumina's Dragen platform (v.3.7.5)[51] against their default graph, which was constructed using the same GRCh38 reference with decoys but no ALTs, and population contigs, SNPs and liftover sequences from datasets internal to their platform. DeepTrio then called small variants from the aligned reads. The same version and parameters were used for DeepTrio (v.1.3). Only the default model was used for these conditions. We also applied the native Dragen caller and joint genotyper to the Dragen-Graph-based alignments for comparison purposes, referred to as Dragen-DragenCall and Dragen-DragenJointCall, respectively. Dragen-DragenCall implements a single-sample based method and is what is the default use-case for processing Dragen-Graph-mapped data. Dragen-DragenJointCall uses a pedigree-backed implementation that informs which variants are likely to be de novo and which are erroneous given the genotype information of the parents. To make a fairer comparison with Dragen, we tested these configurations to assess what implementation of Dragen variant calling produced the best results given the available trio data.

### Evaluation using the GIAB benchmark

The small variant calls were evaluated using HG001–HG007 with the GIAB benchmark (v.4.2.1)[124], on HG002 in challenging medically relevant autosomal genes[47], and on HG002 using a preliminary draft assembly-based benchmark. For the draft assembly-based benchmark, we used Dipcall[38] to align a scaffolded, high-coverage Trio-Hifiasm assembly[21,38] to GRCh38 and call variants, and then we excluded structurally variant regions from the dip.bed file as described above for the benchmarking of small variants from the pangenome graph. The comparison between the call sets and truth set was made with RTG's vcfeval[83] and Illumina's hap.py tool[82] on confident regions of the benchmark. We used high-coverage read sets of the GIAB

HG001, HG002 and HG005 trio child samples and evaluated performance within the held-out chromosome 20 for the GIAB (v.4.2.1) truth set, or the entire genome for the reduced truth set of the challenging medically relevant autosomal genes. The evaluation was also stratified using the set of regions provided by the GIAB at GitHub (https://github.com/genome-in-a-bottle/genome-stratifications)[125].

### Variant calls across samples from the 1KG

We applied our small variant calling pipeline to the high-coverage read sets for the 3,202 samples of the 1KG[19]. The output alignments, in the GAM format, and the VCFs were saved in public buckets at https://console.cloud.google.com/storage/browser/brain-genomics-public/research/cohort/1KGP/vg/graph_to_grch38. We selected 100 trios among those samples to further evaluate the quality of the calls. We tested all variants that have at least one alternative allele in a trio for Mendelian consistency. In addition, for each variant, we only considered trios for which the child's genotype was different from the genotype of at least one of the parents to minimize bias created by systematic calls (for example, all homozygous or all heterozygous). We looked at the fraction of variants–trios that failed Mendelian consistency in the entire genome and in sites that do not overlap simple repeats as defined by the 'simpleRepeat' track downloaded from the UCSC Genome Browser. The results were compared with Mendelian consistency of calls provided by the 1KG that used GATK HaplotypeCaller on the reads aligned to GRCh38. We also repeated this analysis on the two trios of the GIAB (v.4.2.1) benchmark (HG002–HG007) and across the different methods of our evaluation described above (BWA-MEM and DragenGraph mappers; DeepVariant, DeepTrio and Dragen variant callers).

## SV genotyping with PanGenie

### VCF preprocessing

We used a VCF file created on the basis of snarl traversal of the MC graph as a basis for genotyping. The records contained in this VCF represent

bubbles in the underlying pangenome graph and their nested variants, derived from the snarl tree. Each variant was marked according to their level in this tree. Variants annotated by 'LV=0' correspond to the top-level bubbles. We used vcfbub (v.0.1.0)[100] with parameters -l 0 and -r 100000 to filter the VCF. This removed all non-top-level bubbles from the VCF unless they were nested inside a top-level bubble with a reference length exceeding 100 kb; that is, top-level bubbles longer than that are replaced by their child nodes in the snarl tree. The VCF also contained the haplotypes for all 44 assembly samples, representing paths in the pangenome graph. We additionally removed all records for which more than 20% of all 88 haplotypes carried a missing allele ("."). This resulted in a set of 22,133,782 bubbles. In a next step, we used PanGenie (v.1.0.0)[54] to genotype these bubbles across all 3,202 samples from the 1KG based on high-coverage Illumina reads[19].

### Decomposition of variants

We genotyped all top-level bubbles across all 1KG samples. Whereas biallelic bubbles can be easily classified representing SNPs, indels or SVs, this becomes more difficult for multiallelic bubbles contained in the VCF. In particular, larger multiallelic bubbles can contain a high number of nested variant alleles overlapping across haplotypes, represented as a single bubble in the graph. This is especially problematic when comparing the genotypes computed for the entire bubble to external call sets, as coordinates of the bubble do not necessarily represent the exact coordinates of individual variant alleles carried by a sample in such a region (Supplementary Fig. 20).

To tackle this problem, we implemented a decomposition approach that aimed to detect all variant alleles nested inside multiallelic top-level records. The idea was to detect variants from the node traversals of the reference and alternative alleles of all top-level bubbles. Given the node traversals of a reference and alternative path through a bubble, our approach was to match each reference node to its leftmost occurrence in the alternative traversal, resulting in an alignment of the node traversals (Supplementary Fig. 21a). Nested alleles could then be determined based on indels and mismatches in this alignment. As the node traversals of the

alternative alleles can visit the same node more than once (which is not the case for the reference alleles of the MC graph), this approach is not guaranteed to reconstruct the optimal sequence alignment underlying the nodes in these repeated regions.

As an output, the decomposition process generated two VCF files. The first one is a multiallelic VCF that contains exactly the same variant records as the input VCF, just that annotations for all alternative alleles of a record were added to the identifier tag in the INFO field. For each alternative allele, the identifier tag contains identifiers encoding all nested variants it is composed of, separated by a colon. The second VCF is biallelic and contains a separate record for each nested variant identifier defining reference and alternative allele of the respective variant (Supplementary Fig. 21b). Both VCFs are different representations of the same genomic variation, that is, before and after decomposition. We applied this decomposition method to the MC-based VCF file, used the multiallelic output VCF as input for PanGenie to genotype bubbles, and used the biallelic VCF as well as the identifiers to translate PanGenie's genotypes for bubbles to genotypes for all individual nested variant alleles. All downstream analyses of the genotypes are based on this biallelic representation (that is, after decomposition).

Although the majority of short bubbles (<10 bp) are biallelic, particularly large bubbles (>1,000 bp) tend to be multiallelic. Sometimes each of the 88 non-reference (neither GRCh38 nor T2T-CHM13) haplotypes contained in the graph covered a different path through such a bubble (Supplementary Fig. 22a), leading to a VCF record with 88 alternative alleles listed. We determined the number of variant alleles located inside biallelic and multiallelic bubbles in the pangenome after decomposition. As expected, the majority of SV alleles was located inside of the more complex, multiallelic regions of the pangenome (Supplementary Fig. 22b).

## Genotyping evaluation based on assembly samples

We conducted a leave-one-out experiment to evaluate PanGenie's genotyping performance for the call set samples. For this purpose, we repeatedly removed one of the panel samples from the MC VCF and

genotyped it using only the remaining samples as an input panel for PanGenie. We later used the genotypes of the left-out sample as ground truth for evaluation. We repeated this experiment for five of the call set samples (HG00438, HG00733, HG02717, NA20129 and HG03453) using 1KG high-coverage Illumina reads[19]. PanGenie is a re-genotyping method. Therefore, like every other re-typer, it can only genotype variants contained in the input panel VCF, that is, it is not able to detect variants unique to the genotyped sample. For this reason, we removed all variant alleles (after decomposition) unique to the left-out sample contained in the truth set for evaluation. To evaluate the genotype performance, we used the weighted genotype concordance[54]. Extended Data Fig. 9 shows the results stratified by different regions. Extended Data Fig. 9a shows concordances in biallelic and multiallelic regions of the MC VCF. The biallelic regions include only bubbles with two branches. The multiallelic regions include all bubbles in which haplotypes cover more than two different paths. Extended Data Fig. 9b shows the same results stratified by genomic regions defined by GIAB that we obtained from the following genotypes: easy (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/union/GRCh38_notinalldifficultregions.bed.gz); low-mappability (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/union/GRCh38_alllowmapandsegdupregions.bed.gz); repeats (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/LowComplexity/GRCh38_AllTandemRepeats_gt100bp_slop5.bed.gz); and other-difficult (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/OtherDifficult/GRCh38_allOtherDifficultregions.bed.gz).

Here and in the following, we considered results for SNPs, indels (1–49 bp), SV deletions, SV insertions and other SV alleles, defined as follows: SV deletions include all alleles for which length(REF) ≥50 bp and length(ALT) = 1 bp; SV insertions include all alleles for which length(REF) = 1 bp and length(ALT) ≥ 50 bp; all other alleles with a length ≥50 bp are included in ‘others’.

Overall, weighted genotype concordances were high for all variant types. In particular, variant alleles in biallelic regions of the graph were easily genotypable. Alleles inside multiallelic bubbles were more difficult to genotype correctly as PanGenie needs to decide between several possible alternative paths, whereas there are only two such paths for biallelic regions (Extended Data Fig. 9a). Furthermore, genotyping accuracy depended on the genomic context (Extended Data Fig. 9b). Regions with low mappability, repetitive regions and other difficult regions were harder to genotype than regions classified as 'easy' by GIAB.

## Creating a high-quality subset

We generated genotypes for all 3,202 1KG samples with PanGenie and defined a high-quality subset of SV alleles that we could reliably genotype. For this purpose, we applied a machine-learning approach similar to what we have previously presented[5,54]. We defined positive and negative subsets of variants based on the following filters: ac0_fail, a variant allele was genotyped with an AF of 0.0 across all samples; mendel_fail, the mendelian consistency across trios is <80% for a variant allele. Here, we use a strict definition of Mendelian consistency, which excludes all trios with only 0/0, only 0/1 and only 1/1 genotypes; gq_fail, <50 high-quality genotypes were reported for this variant allele; self_fail, genotyping accuracy of a variant allele across the panel samples is <90%; nonref_fail, not a single non-0/0 genotype was genotyped correctly across all panel samples.

The positive set included all variant alleles that passed all five filters. The negative set contained all variant alleles that passed the ac0_fail filter but failed at least three of the other filters. We trained a support vector regression approach based on these two sets that used multiple features, including AFs, Mendelian consistencies or the number of alternative alleles transmitted from parents to children. We applied this method to all remaining variant alleles genotyped with an AF > 0, resulting in a score between –1 (bad) and 1 (good) for each. We finally defined a filtered set of variants that included the positive set and all variant alleles with a score of ≥ –0.5.

We show the number of variant alleles contained in the unfiltered set, the positive set and the filtered set in Supplementary Table 18. As our focus is on SVs and as 65% of all SNPs and indels are already contained in the positive set, we applied our machine-learning approach only to SVs. We found that 50%, 33% and 26% of all deletion, insertion and other alleles, respectively, were contained in the final, filtered set of variants. Note that these numbers take all distinct SV alleles contained in the call sets into account. Especially for insertions and other SVs, many of these alleles are highly similar, with sometimes only a single base pair differing. Therefore, it is probable that many of these actually represent the same events. Our genotyping and filtering approach helped to remove such redundant alleles.

To evaluate the quality of the PanGenie genotypes, we compared the AFs observed for the SV alleles across all 2,504 unrelated 1KG samples to their AFs observed across the 44 assembly samples in the MC call set. Supplementary Figs. 23–25 show the results for SV deletions, insertions and other SV alleles. We observed that the AFs between both sets matched well, resulting in Pearson correlations of 0.93, 0.87 and 0.81 for deletions, insertions and other alleles, respectively, contained in the unfiltered set. For the filtered set, we observed correlations of 0.96, 0.93 and 0.90, respectively. We also analysed the heterozygosity of the PanGenie genotypes across all 2,504 unrelated 1KG samples and observed a relationship close to what is expected by Hardy–Weinberg equilibrium (Supplementary Figs. 23–25, lower panels).

### Number of SVs per sample

We compared our filtered set of variant alleles to the HGSVC PanGenie genotypes (v.2.0 lenient set)[5] and Illumina-based SV genotypes[19]. A direct comparison of the three call sets is difficult. The HGSVC and HPRC call sets are based on variant calls produced from haplotype-resolved assemblies of 32 and 44 samples, respectively[5]. For each call set, variants were re-genotyped across all 3,202 1KG samples. Note that the call set samples for HPRC and HGSVC are disjoint. As re-genotyping cannot discover new variants, both call sets will miss variants carried by 3,202 samples that were not seen in the assembly samples. By contrast, the 1KG call set contains short-read based variant calls produced for each of the 3,202 1KG samples.

Another difference between the HGSVC and HPRC call sets is that in the HGSVC call set, highly similar alleles are merged into a single record to correct for representation differences across different samples or haplotypes. The HPRC call set, however, keeps all these alleles separately even if there is only a single base pair difference between them. To make the call sets better comparable, we merged clusters of highly similar alleles in the HPRC filtered set before comparisons with other call sets. This was done with Truvari (v.3.1.0)[112] using the command: truvari collapse -r 500 -p 0.95 -P 0.95 -s 50 -S 100000.

To be able to properly compare the call sets despite their differences, we counted the number of SV alleles present in each sample (genotype 0/1 or 1/1) in each call set and plotted the corresponding distributions stratified by genome annotations from GIAB (same as above, Fig. 6d). We also generated the same plot including only common SV alleles with an AF > 5% across all 3,202 samples (Supplementary Fig. 28). Both plots showed that both assembly based call sets (HPRC and HGSVC) were able to access more SVs across the genome than the short-read-based 1KG call set, especially deletions <300 bp and insertions (Fig. 6e). This result confirms that SV callers based on short reads alone miss a large proportion of SVs located in regions inaccessible by short-read alignments, which has been previously reposed by several studies[5,8]. In the 'easy' regions, the number of SVs per sample was consistent across all three call sets. For the other regions, however, results indicated that the HPRC-filtered genotypes provided access to more variant alleles than the HGSVC lenient set, especially insertions and variants in regions of low mappability and tandem repeats (Fig. 6d,e).

To evaluate the new SVs in our filtered HPRC call set, we revisited the leave-one-out experiment we had previously performed on the unfiltered set of variants (see above). We restricted the evaluation to the following subset of variants: (1) those that are in our filtered set but not in the 1KG Illumina calls (novel); (2) those in our filtered set and in the 1KG Illumina call set (known); and (3) all variants in our filtered set. To find matches between our set and the Illumina calls, we used a criterion based on reciprocal overlap of at least 50%. Results are shown in Supplementary Fig. 29. We generated two versions of this figure: the first one (top) excludes variants

that are unique to the left-out sample and therefore not typable by any re-genotyping method, and the second one includes these variants (bottom). In general, genotype concordances of all lenient variants (brown, dark purple) were slightly higher compared with the concordances we observed for the unfiltered set (Supplementary Fig. 29). Furthermore, concordances of the known variants were highest. This is expected, as these variants tended to be in regions easier to access by short reads. Concordances for novel variants were slightly worse. This was also expected, as these variants tended to be located in more complex genomic regions that are generally harder to access. However, even for these variants, concordances were still high, which indicated that the PanGenie genotypes for these variants are of high quality.

### Evaluation based on medically relevant SVs

In addition to all 3,202 1KG samples, we genotyped sample HG002 based on Illumina reads from ref. [18]. We used the GIAB CMRG benchmark containing medically relevant SVs[47] downloaded from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/GRCh38/StructuralVariant/ for evaluation. Similar to the 1KG samples, we used the MC-based VCF (see above) containing variant bubbles and haplotypes of 44 assembly samples as an input panel for PanGenie. We extracted all variant alleles with a length ≥50 bp from our genotyped VCF (biallelic version, after decomposition). We converted the ground truth VCF into a biallelic representation using bcftools norm -m -any and kept all alleles with length ≥50 bp. We used Truvari (v.3.1.0)[112] with parameters --multimatch --includebed <medically-relevant-sv-bed> -r 2000 --no-ref a -C 2000 --passonly to compare our genotype predictions to the medically relevant SVs. Results are shown in Supplementary Table 22. As PanGenie is a re-typing method, it can only genotype variants provided in the input and therefore cannot detect novel alleles. As HG002 is not among the panel samples, the input VCF misses variants unique to this HG002 sample. Thus, these unique variants cannot be genotyped by PanGenie and will be counted as false negatives during evaluation. Therefore, we computed an adjusted version of the recall that excluded SV alleles unique to HG002 (that is,

alleles not in the graph) from the truth set for evaluation. To identify which SV alleles were unique, we compared each of the 44 panel samples to the ground truth VCF using Truvari to identify the false negatives for each sample. Then we computed the intersection of false-negative calls across all samples. The resulting set then contained all variant alleles unique to the HG002 ground truth set. We found 15 such unique SV alleles among the GIAB CMRG variants. We removed these alleles from the ground truth set and recomputed precision–recall statistics for our genotypes. Adjusted precision–recall values are shown in Supplementary Table 22.

All genotyping results produced by PanGenie are available at Zenodo (https://doi.org/10.5281/zenodo.6797328).

## Read mapping at VNTR regions

### Simulating and mapping VNTR reads

Raw VNTR coordinates on GRCh38 (chromosomes 1–22 and sex chromosomes only) were generated using TRF (v.4.09)[56] with command trf hg38.fa 2 7 7 80 10 50 500 -f -d. Only repeats with a period size between 6,000 and 10,000 bp, total length >100 bp and not overlapping with centromeric regions were selected, leaving a total number of 98,021 non-overlapping loci. Using the raw VNTR coordinates on GRCh38 as input, VNTR regions across 96 haplotypes (including GRCh38 and CHM13) were annotated using the build module in danbing-tk (v.1.3)[55] (dist_scan=700, dist_merge=1, TRwindow=100000, MBE_th1=0.3, MBE_th2=0.6).

Whole-genome paired-end error-free short reads were simulated at around 30× for each genome, or equivalently about 15× for each haplotype. A read pair was generated for every 20 bp with fragment size of 500 bp and read length of 150 bp. Paired-end read mapping to the MC graph was done using Giraffe (v.1.39.0)[49] using the command vg giraffe -x $pref.xg -g $pref.gg -H $pref.gbwt -m $pref.min -d $pref.dist -p -f <(zcat $h1 $h2) -i -t 16, whereas mapping to GRCh38 was done using BWA-MEM (v.0.7.17-r1188)[50] using the command bwa mem -t 16 -Y -K 100000000 -p $ref <(zcat $h1 $h2). For a fair comparison, GRCh38 plus decoy minus

ALT/HLA contigs were used as reference to match the paths included in the MC graph.

## Evaluating read mapping accuracy at VNTR regions

To evaluate the performance of read mapping using the MC graph plus Giraffe, the VNTR information from danbing-tk were used to annotate each node in the graph by traversing each haplotype path. Every node that covers a VNTR region has a tuple that denotes the intersected interval; any aligned reads overlapping with the interval were considered mapped to the VNTR. Similarly, a read simulated from an interval overlapping with a VNTR was considered derived from the VNTR. To evaluate the performance of GRCh38 plus BWA-MEM, the mapped region by each read was obtained using the bamtobed submodule in BEDTools (v.2.30.0)[123]. The VNTR annotations on GRCh38 were used to determine whether a read was mapped to a VNTR.

For each read, we tracked its source and mapped VNTR and VNTRs, and used this information to compute accuracy. Only VNTRs present in danbing-tk's annotations were tracked; otherwise they were labelled untracked, the same as non-VNTR regions. A true positive denotes mapping from a VNTR to its original VNTR. An exogenous false positive denotes mapping from untracked regions to a VNTR. An endogenous false positive denotes mapping from a VNTR to another VNTR. A true negative denotes mapping from untracked regions to untracked regions. A false negative denotes mapping from a VNTR to untracked regions (Supplementary Fig. 30). Any alignments in the JSON output of Giraffe that did not contain the mapping field were considered unmapped. The two ends of a read pair that did not map to the same chromosome by BWA-MEM were also considered unmapped.

## Estimating VNTR length variants from read depths

The WGS samples for 35 genomes (HG00438, HG00621, HG00673, HG00733, HG00735, HG01071, HG01106, HG01109, HG01175, HG01243, HG01258, HG01361, HG01891, HG01928, HG01952,

HG01978, HG02055, HG02080, HG02145, HG02148, HG02257, HG02572, HG02622, HG02630, HG02717, HG02723, HG02818, HG02886, HG03098, HG03453, HG03486, HG03492, HG03579, NA18906 and NA19240) were mapped to the MC graph using Giraffe as described in the 'Pangenome point genotyping' subsection. Using the VNTR annotations described in the previous section, the number of reads mapped to each VNTR region in the MC graph was calculated as a proxy for VNTR length. VNTRs with invariant length across the 35 genomes were removed from analysis, leaving a total of 60,861 loci. The ground truth for a VNTR in a genome was computed from the number of bases spanned by the VNTR, averaged from the two haplotypes.

As a baseline control, the read depth of each VNTR region for the 35 WGS samples produced by mapping reads to GRCh38 was also computed with mosdepth (v.0.3.1)[126] using the command mosdepth -t 4 -b $VNTR_bed -x -f $hg38 $pref $cram. To be able to compare with the graph-based approach, VNTRs with missing annotation on GRCh38 were further removed, leaving a total of 60,386 VNTRs.

## RNA-seq mapping evaluation

We augmented the allele-filtered graph (see the 'Pangenome point genotyping' subsection) with edges for splice junctions to create a spliced pangenome graph using the rna subcommand in the vg toolkit (v.1.38.0) with a maximum node length set to 32 (vg rna -k 32)[57]. The transcript annotations that were used to define the splice junctions consisted of the CAT transcript annotations on each assembly together with splice junctions from the GENCODE (v.38) annotation[29]. Transcripts from the GENCODE (v.38) annotation were further added as paths to the spliced pangenome graph. For comparison, we also created a spliced reference constructed from the reference sequence, once again using the GENCODE (v.38) transcript annotation. For both graphs, we created the indices needed for mapping using the vg toolkit (v.1.38.0) with default parameters, except when pruning, for which edges on embedded paths were restored (vg prune -r). Furthermore, for the spliced HPRC pangenome graph, it was necessary to use stricter pruning parameters (vg prune -r -k 64 -M 64). For the spliced

reference, we also created the index needed by the RNA-seq mapper STAR using default parameters.

We simulated RNA-seq reads with a pipeline that was designed to preserve complex genome variation in the simulated data. The transcript sequences used for the simulation were derived from the GENCODE (v.38) transcript annotations projected onto assembled haplotypes from HG002. Specifically, we used MC to create an alignment between GRCh38 and the two HPRC HG002 assembly haplotypes, which were held out of the main pangenome graph for benchmarking. We then used CAT to lift the transcript annotations over to these haplotypes. We constructed a spliced personal genome graph using the vg rna subcommand, and then we simulated reads using vg sim (commit 2cea1e2) using an Illumina NovaSeq cDNA read set (SRR18109271) to fit model parameters. This essentially amounts to simulating directly from the projected transcript sequences. The transcripts were simulated with uniform expression, split evenly between the two haplotypes, keeping the reads from each haplotype separate. This expression profile is not biologically realistic, but it avoids the difficulty of choosing a particular expression profile as representative for all tissues and stages of development. Moreover, existing estimated profiles would be biased towards the tools that were used to estimate them. We simulated 5,000,000 paired-end 150 bp RNA-seq reads.

Both simulated and real Illumina RNA-seq reads were mapped to the graphs using vg mpmap (commit c0c4816) with default parameters. In addition, the reads were mapped to the spliced reference using STAR (v.2.7.10a) with default parameters[58]. For the real data, we used previously published data NA12878 RNA-seq data (SRR1153470)[127] and the ENCODE project (ENCSR000AED, replicate 1)[128,129].

We used the same approach as previously described[57] to evaluate the alignments. In brief, for the simulated data, the graph alignments were compared with the truth alignments by estimating their overlap on the reference genome paths. The graph alignments were projected to the reference paths using vg surject -S. A uniquely aligned read (one with primary MAPQ $\geq$ 30) was considered correct if it overlapped 90% of the truth alignment. A multi-mapped read was considered correct if any of the

multi-mappings was correct under the same criterion. For the real data, the average read coverage of each exon on the reference path calculated from the projected graph alignments were compared with the corresponding coverages estimated from long-read alignments. For the long-read data, we used PacBio Iso-Seq alignments from the ENCODE project (ENCSR706ANY, all replicates), which come from the same cell line as the Illumina data. The long-read alignments were used to define the exons, and only primary long-read alignments with a mapping quality of at least 30 were used. The alignments for the four Iso-Seq replicates (ENCFF247TLH, ENCFF431IOE, ENCFF520MMC and ENCFF626GWM) were combined and filtered using SAMtools (v.1.15)[130]. BEDTools (v.2.30.0) was used to convert the alignments to exons coordinates[123].

We also used the results of the mapping experiment to quantify allelic bias. We used vg deconstruct to identify sites of variation in the MC graph of haplotypes from HG002, with deletions greater than 10 kb removed to avoid spurious variants. Variants overlapping exons were selected and normalized using BCFtools (v.1.16)[130]. Next we counted the number of mapped reads from each of the two haplotypes that overlapped each heterozygous exonic variant[57]. Specifically, the read count for each allele was calculated as the average count across the two breakpoints of an allele. This was done to treat different variant types and lengths equally. We then tested all variants with a read coverage of at least 20 for allelic bias using a two-sided binomial test at $\alpha = 0.01$. All tests that reject the null hypothesis are false positives, as the reads were simulated without allelic bias. The results were split into different classes of variants and plotted against the number of sites that achieved a coverage of at least 20, which is a rough indicator of mapping sensitivity. Indels larger than 50 bp were excluded.

We also compared gene expression inference using the mapped reads. For the vg mpmap (v.1.43.0) graph mappings, we used RPVG (commit 1d91a9e)[57] in transcript inference mode to quantify expression. Two different transcript annotations were used as input to RPVG. The MC pantranscriptome was created from the CAT transcript annotations on each assembly and the GENCODE (v.38) transcriptome. The mpmap-RPVG pipeline was compared with Salmon (v.1.9.0)[131] and RSEM (v.1.3.3)[132]. These methods were provided the GENCODE (v.38) transcriptome as input,

both with and without transcripts on the GRCh38 alternative contigs. Any genes unique to the alternative contigs were filtered. For Salmon, the GRCh38 reference was used as a decoy and duplicate transcripts were kept. Bowtie2 (v.2.4.5)[133] was used as a mapper for RSEM. Gene expression values were calculated by summing the corresponding transcript level values for each gene. GffRead (v.0.12.7) was used to create a table of gene names and transcript identifiers from the transcript annotation[134]. The accuracy was measured in two ways: (1) the Spearman correlation between the simulated and inferred expression values, and (2) the mean absolute relative difference between the simulated and inferred expression values.

The scripts that were used for graph construction, read simulation, mapping and evaluation are available at GitHub (https://github.com/jonassibbesen/hprc-rnaseq-analyses-scripts).

## ChIP-seq analysis

We aligned H3K4me1, H3K27ac and ATAC-seq obtained from monocyte-derived macrophages from 30 individuals[59] using vg map[13] to the GRCh38 reference genome graph and to the HPRC genome graph. Then we called peaks using Graph Peak Caller (v.1.2.3)[135] on both sets of alignments for each of the 30 H3K4me1, H3K27ac and ATAC-seq samples. To identify HPRC-only peaks, we projected HPRC coordinates to the GRCh38 path using Graph Peak Caller and compared intervals using BEDTools[123]. We named HPRC peaks that overlapped GRCh38 peaks as common peaks and those that do not as HPRC-only. We calculated the expected frequency (as inverse cumulative distributions) of common and HPRC-only peaks among the 30 samples by resampling the peaks of each sample from the peaks of all the samples and re-counted the number of overlaps. We repeated this simulation 100 times and plotted the average curves. We determined heterozygous variants in our samples by aligning WGS datasets for each sample to the HPRC graph using vg map and genotyping the variants using vg call -a. We narrowed the list of heterozygous SVs above 50 bp in each sample with the aim of looking for allelic-specific peaks. For each epigenomic sample, we obtained allelic-specific read counts within peaks that lie on the previously identified loci by running vg call -a on the epigenomic HPRC alignments, which outputs the numbers of reads on each

path in a bubble (DP and AD fields in the VCF output). We then assigned peaks to the SV or reference allele, or both alleles with a two-tailed binomial test parameterized on the sum of reads on both alleles and $P$ = 0.05. Any peak with reads on one allele, but not on the other was assigned to the allele with the reads. Read counts were proportionally adjusted for the difference in length between the reference and SV alleles.

Processed data, scripts and code for the above steps are available at Zenodo (https://doi.org/10.5281/zenodo.6564396).

## Population genetic analyses

Although the size of the population sample represented in our pangenome is small, it provides access to previously under-ascertained regions of the genome. We sought to understand the potential utility of these regions for future population genetic studies using regional PCA based on variants called compared to the CHM13 and GRCh38 references. For these analyses, we considered both the PGGB (whole pangenome, combined) and MC (reference-based, distinct CHM13 and GRCh38) graphs. For both graph models, the CHM13 VCFs provide access to regions that were not previously observed by studies based on GRCh38, for which short-read-based studies may have difficulty reliably aligning and calling variants. In combination, these two graphs provide cross-validation of implied population genetic patterns in these new regions, which we explore here.

To understand chromosome-specific patterns of variation, we applied PCA to each autosomal chromosome independently to the VCFs from PGGB (PGGB-CHM13, PGGB-GRCh38).

To ensure that observed patterns were not derived from higher rates of assembly error in the repetitive regions of acrocentric p-arms, we used our Flagger-confident region annotations to prune the PGGB graph (using odgi inject to inject the confident regions as subpaths and then odgi prune to remove the full original paths that were including unreliable regions) to only confident regions of assemblies. We then reapplied vg deconstruct to this graph to obtain a new set of SNPs (the code for the PGGB graphs pruning and variant calling on the pruned graphs can be found at the

following link:
https://github.com/pangenome/HPRCyear1v2genbank/blob/main/workflows/confident_variants.md). Genome-wide, we found that pruning reduced the number of called SNPs by only 1.188% (previous $N = 23{,}272{,}652$, pruned $N = 22{,}996{,}113$). The total reduction in the acrocentrics was higher, with 6.29% fewer SNPs (previous $N = 3{,}735{,}605$, pruned $N = 3{,}676{,}746$), which indicated the difficulty in assembling these regions. We note that the PCA sample distributions remained almost identical (data not shown), which indicated that the patterns observed in the full graph are maintained despite assembly issues. In these filtered PGGB-CHM13 and PGGB-GRCh38 VCFs, we considered all biallelic SNPs relative to the chosen reference, regardless of variant nesting level (data not shown; filtering for only SNPs LV = 0 or LV > 0 produced nearly identical results). A qualitative evaluation suggested no significant differences in PCA patterns across the metacentric chromosomes (Supplementary Fig. 47). However, in the p-arms of the acrocentrics (chromosomes 13, 14, 15, 21 and 22), which are accessible in the PGGB-CHM13 VCF, we observed a reduction in population differentiation and a higher rate of variance explained in the lowest principal component.

To investigate this quantitatively, we measured the number of clusters implied by the PCA for the PGGB-CHM13 VCFs using *k*-means clustering to automatically determine the optimal number of clusters for each PCA (gap_stat clustering in the fviz_nbclust function of the factoextra R package) (analysis code available at:
https://github.com/SilviaBuonaiuto/hprcPopGenAnalysis). Applying this approach to three PCAs per chromosome VCF, we obtained optimal cluster counts for the p-arm, q-arm and entire chromosome. In metacentric chromosomes, we usually observed optimal numbers of clusters approximately corresponding to the number of expected world population groupings in the input genomes (3–5, as in Supplementary Fig. 48). However, in the p-arms of the acrocentrics, we observed much fewer, in general only one cluster, indicative of reduced population differentiation compared with other parts of the acrocentric chromosomes. This pattern was only apparent in the PGGB graph based on CHM13. To evaluate the difference quantitatively, we applied a Wilcoxon rank-sum test to compare the differences between cluster count distributions in metacentric versus

acrocentric chromosomes across the entire chromosome, the q-arm and the p-arm. There were non-significant differences between the distributions between acrocentric and metacentric chromosomes at a chromosome scale and in the q-arms, but a significant difference (Wilcoxon $P = 0.013$) in the case of acrocentric p-arms (Supplementary Fig. 49).

This analysis indicates that significant challenges remain for the use of these new regions in population genetic studies. Patterns observed in PCA projections of the pangenome across all chromosomes suggests a distinct process of variation sharing between populations within the short arms of the acrocentrics. In effect, we observed a more homogenous population in these regions when using the CHM13 assembly as a reference. This reference contains real sequences in these regions, whereas GRCh38 contains gaps, which render analysis impossible. The apparent population homogenization could be driven by error. We mitigated this issue by utilizing only SNPs found in Flagger-confident regions, but this does not guard against potential sources of alignment error that are likely to be amplified by the repetitive sequences in these loci. It is also possible that the chromosome-specific partitioning process applied by both graph models is failing to correctly partition contigs on these short arms. The known homology between the short arms bolsters the possibility of ongoing sequence information exchange between non-homologous chromosomes[3], which would be consistent with the patterns we observed. In summary, this analysis shows that when using CHM13 as a reference, the behaviour of sequences on the short arms of the acrocentrics in the PGGB graph is not similar to that of other sequences in the pangenome.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Sequencing data, assemblies and pangenomes produced by the HPRC are available at AnVIL (https://anvilproject.org) in the AnVIL_HPRC

workspace. Data are also available as part of the AWS Open Data Program (https://registry.opendata.aws) in the human-pangenomics S3 bucket (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html). In addition, data have been uploaded to the International Nucleotide Sequence Database Collaboration for long-term storage and availability. Supporting information about the data (including index files with S3 and GCP file locations) can be found in our GitHub repositories (see below). Sequencing data for 29 selected HPRC samples from the 1KG cohort (Results) are uploaded to BioProject PRJNA701308 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA701308). Sequencing data created by the HPRC for samples in the cohort of 18 additional samples are uploaded to BioProject PRJNA731524 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA731524). Both sets of assemblies are grouped by an Umbrella BioProject PRJNA730822. Data used in this paper have additional information available at AnVIL and at GitHub (https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0). Assemblies along with assembly annotations, such as RepeatMasker and Ensembl gene annotations, can be viewed in an assembly hub in the UCSC Genome Browser (http://hprc-browser.ucsc.edu). The data and annotations can also be accessed through the Ensembl Rapid Release Genome Browser (https://rapid.ensembl.org) and a dedicated Ensembl project page for centralized access to HPRC data (https://projects.ensembl.org/hprc), which includes links to download the data locally. File locations for the assemblies (and select annotation files) that are stored in S3 and GCP can also be found at AnVIL or GitHub (https://github.com/human-pangenomics/HPP_Year1_Assemblies). Pangenomes were uploaded to the European Nucleotide Archive as analysis objects and are organized under Umbrella BioProject PRJNA850430 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA850430). File locations for pangenomes, indices and variant calls derived from the pangenomes can be found at AnVIL or GitHub (https://github.com/human-pangenomics/hpp_pangenome_resources). Variant calls produced by the analysis performed in this paper are available at Amazon S3 (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=publications/PANGENOME_2022) or where indicated in the relevant sections below. PanGenie genotypes produced for the 1KG

samples based on the MC graph are available at Zenodo (https://doi.org/10.5281/zenodo.6797328).

# Code availability

The list of all tools, including versions and/or code commits, used for this study are available in Supplementary Table 20.

# References


1. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

3. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

4. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

5. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

6. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).

7. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

8. Zhao, X. et al. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome

sequencing technologies. *Am. J. Hum. Genet.* **108**, 919–928 (2021).

9. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).

10. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

11. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2018).

12. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).

13. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).

14. Eizenga, J. M. et al. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).

15. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).

16. Shafin, K. et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).

17. Porubsky, D. et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021).

18. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).

19. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).

20. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

21. Jarvis, E. D. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).

22. Guarracino, A. et al. Recombination between heterologous human acrocentric chromosomes. *Nature* https://doi.org/10.1038/s41586-023-05976-y (2023).

23. Garg, S. Computational methods for chromosome-scale haplotype reconstruction. *Genome Biol.* **22**, 101 (2021).

24. Garg, S. Pstools: a toolkit for fully phased sequences on chromosome level. *GitHub* https://github.com/shilpagarg/pstools (2020).

25. Rahman, A. & Pachter, L. CGAL: computing genome assembly likelihoods. *Genome Biol.* **14**, R8 (2013).

26. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0, http://www.repeatmasker.org (2013–2015).

27. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).

28. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).

29. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).

30. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).

31. Sudmant, P. H. et al. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).

32. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).

33. Porubsky, D. et al. Gaps and complex structurally variant loci in phased genome assemblies. *Genome Res.* https://doi.org/10.1101/gr.277334.122 (2023).

34. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).

35. Hickey, G. et al. Pangenome graph construction from genome alignment with Minigraph-Cactus. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.06.511217 (2022).

36. Garrison, E. et al. Building pangenome graphs. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2023.04.05.535718v1 (2023).

37. Armstrong, J. et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).

38. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).

39. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).

40. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

41. Pacific Biosciences. PBSV: a suite of tools to call and analyze structural variants in diploid genomes from PacBio SMRT reads. *GitHub* https://github.com/PacificBiosciences/pbsv (2021).

42. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–

468 (2018).

43. Kirsche, M. et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 (2023).

44. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).

45. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2020).

46. Fiddes, I. T. et al. Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. *Genome Res.* **28**, 1029–1038 (2018).

47. Wagner, J. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).

48. Chen, X. et al. Cyrius: accurate *CYP2D6* genotyping using whole-genome sequencing data. *Pharmacogenomics J.* **21**, 251–261 (2021).

49. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).

50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1303.3997 (2013).

51. Miller, N. A. et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med.* **7**, 100 (2015).

52. Kolesnikov, A. et al. DeepTrio: variant calling in families using deep learning. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.05.438434 (2021).

53. Vollger, M. R. et al. Increased mutation and gene conversion within human segmental duplications. *Nature* https://doi.org/10.1038/s41586-023-05895-y (2023).

54. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).

55. Lu, T.-Y., Human Genome Structural Variation Consortium & Chaisson, M. J. P. Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nat. Commun.* **12**, 4250 (2021).

56. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

57. Sibbesen, J. A. et al. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nat. Methods* **20**, 239–247 (2023).

58. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

59. Groza, C. et al. Genome graphs detect human polymorphisms in active epigenomic state during influenza infection. Preprint at *bioRxiv* https://doi.org/10.1101/2021.09.29.462206 (2022).

60. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e19 (2019).

61. Falchi, M. et al. Low copy number of the salivary amylase gene predisposes to obesity. *Nat. Genet.* **46**, 492–497 (2014).

62. Mohajeri, K. et al. Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the chromosome 8p23.1 region. *Genome Res.* **26**, 1453–1467 (2016).

63. Fiddes, I. T. et al. Human-specific *NOTCH2NL* genes affect Notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369.e22 (2018).

64. Eizenga, J. M. et al. Efficient dynamic variation graphs. *Bioinformatics* **36**, 5139–5144 (2021).

65. Sirén, J. & Paten, B. GBZ file format for pangenome graphs. *Bioinformatics* **38**, 5012–5018 (2022).

66. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).

67. Groza, C., Kwan, T., Soranzo, N., Pastinen, T. & Bourque, G. Personalized and graph genomes reveal missing signal in epigenomic data. *Genome Biol.* **21**, 124 (2020).

68. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).

69. Tang, D. et al. Genome evolution and diversity of wild and cultivated potatoes. *Nature* **606**, 535–541 (2022).

70. Zhou, Y. et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).

71. Schatz, M. C. et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genom.* **2**, 100085 (2022).

72. Sim, S. HiFiAdapterFilt: remove CCS reads with remnant PacBio adapter sequences and convert outputs to a compressed .fastq (.fastq.gz). *GitHub* https://github.com/sheinasim/HiFiAdapterFilt (2021).

73. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the Integrative Genomics Viewer. *Cancer Res.* **77**, e31–e34 (2017).

74. Li, H. Yak: yet another *k*-mer analyzer. *GitHub* https://github.com/lh3/yak (2020).

75. Li, H. Gfatools: tools for manipulating sequence graphs in the GFA and rGFA formats. *GitHub* https://github.com/lh3/gfatools (2021).

76. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).

77. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

78. National Center for Biotechnology Information. VecScreen: screen a sequence for vector contamination. *Natl Cent. Biotech. Inf.* https://www.ncbi.nlm.nih.gov/tools/vecscreen (2016).

79. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).

80. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

81. Chin, C.-S. et al. A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nat. Commun.* **11**, 4794 (2020).

82. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).

83. Cleary, J. G. et al. Comparing Variant Call Files for performance benchmarking of next-generation sequencing variant calling pipelines. Preprint at *bioRxiv* https://doi.org/10.1101/023754 (2015).

84. Zook, J. Genome in a Bottle—v3.0 genome stratifications. *Natl Inst. Stand. Technol.* https://doi.org/10.18434/mds2-2499 (2021).

85. Numanagic, I. et al. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **34**, i706–i714 (2018).

86. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

87. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

88. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2020).

89. Lee, C., Grasso, C. & Sharlow, M. F. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**, 452–464 (2002).

90. Li, H. Identifying centromeric satellites with dna-brnn. *Bioinformatics* **35**, 4408–4410 (2019).

91. Gao, Y. et al. abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. *Bioinformatics* **37**, 2209–2211 (2021).

92. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).

93. Hickey, G. Hal2vg: convert HAL to vg-compatible sequence graph. *GitHub* https://github.com/ComparativeGenomicsToolkit/hal2vg (2021).

94. Doerr, D. & Marijon P. GFAffix: identify walk-preserving shared affixes in variation graphs and collapses them into a non-redundant graph structure. *GitHub* https://github.com/marschall-lab/gfaffix (2022).

95. Guarracino, A., Mwaniki, N., Marco-Sola, S. & Garrison, E. wfmash: a pangenome-scale aligner. *Zenodo* https://doi.org/10.5281/zenodo.6949373 (2021).

96. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *Bioinformatics* **39**, btac743 (2023).

97. Garrison, E. et al. smoothxg: normalization of variation graphs with local partial order realignment. *Zenodo* https://doi.org/10.5281/zenodo.7239231 (2022).

98. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).

99. Paten, B. et al. Superbubbles, ultrabubbles, and cacti. *J. Comput. Biol.* **25**, 649–663 (2018).

100. Garrison, E. vcfbub: popping bubbles in vg deconstruct VCFs. *Zenodo* https://doi.org/10.5281/zenodo.7239225 (2022).

101. Li, H. SDUST: symmetric DUST for finding low-complexity regions in DNA sequences. *GitHub* https://github.com/lh3/sdust (2019).

102. Li, H. ETRF: exact tandem repeat finder. *GitHub* https://github.com/lh3/etrf (2019).

103. Li, H. Seqtk: a toolkit for processing sequences in FASTA/Q formats. *GitHub* https://github.com/lh3/seqtk (2018).

104. Doerr, D. Panacus: calculate growth statistics for pangenome graphs. *GitHub* https://github.com/marschall-lab/panacus (2022).

105. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

106. Pedersen, B. S. et al. Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genom. Med.* **6**, 60 (2021).

107. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* **19**, 705–710 (2022).

108. Martin, M. et al. WhatsHap: fast and accurate read-based phasing. Preprint at *bioRxiv* https://doi.org/10.1101/085050 (2016).

109. Larson, D. E. et al. svtools: population-scale analysis of structural variation. *Bioinformatics* **35**, 4782–4787 (2019).

110. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput. Biol.* **18**, e1009123 (2022).

111. Marco-Sola, S. et al. Optimal gap-affine alignment in *O(s)* space. *Bioinformatics* **39**, btad074 (2023).

112. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).

113. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).

114. Hickey, G. et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).

115. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).

116. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).

117. Lin, M. F. Gfabase: GFA insert into GenomicSQLite. *GitHub* https://github.com/mlin/gfabase (2021).

118. Wilkins, D. Gggenes: a ggplot2 extension for drawing gene arrow maps. *GitHub* https://github.com/wilkox/gggenes (2022).

119. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

120. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1207.3907 (2012).

121. Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M. & Parker, J. S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* **30**, 2813–2815 (2014).

122. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at *bioRxiv* https://doi.org/10.1101/201178 (2017).

123. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

124. Wagner, J. et al. Benchmarking challenging small variants with linked and long reads. *Cell Genom.* **2**, 100128 (2022).

125. Olson, N. D. et al. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom.* **2**, 100129 (2022).

126. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).

127. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl Acad. Sci. USA* **111**, 9869–9874 (2014).

128. Davis, C. A. et al. The Encyclopedia of DNA Elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

129. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

130. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

131. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

132. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

133. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

134. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res.* **9**, 304 (2020).

135. Grytten, I. et al. Graph Peak Caller: calling ChIP-seq peaks on graph-based reference genomes. *PLoS Comput. Biol.* **15**, e1006731 (2019).

# Acknowledgements

We would like to acknowledge S. Bidwell and other members of the GenBank staff at the National Center for Biotechnology Information (NCBI; NLM/NIH) for their work to release the assemblies into GenBank. Certain commercial equipment, instruments or materials are identified to adequately specify experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the equipment, instruments or materials identified are necessarily the best available for the purpose. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf. This work was funded in part by the National Human Genome Research Institute of the National Institutes of Health under award numbers U41HG010972, 1U01HG010973, U41HG007234, 1R01HG011274, R01HG010485, U24HG010262 U01HG010963, U24HG007497 and R01HG011649. This work was funded in part by the National Institutes of Health under award numbers U01HG010961, OT2OD033761, U24HG011853, R01-HG006677, R35-GM130151, R01HG002385, R01HG010169, U01HG01973, 5U01HG010971, R01GM123489, U24HG009081 and 1ZIAHG200398. This work was funded in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. The work of F.T.-N. and V.A.S. was supported by the National Center for Biotechnology

Information of the National Library of Medicine (NLM), National Institutes of Health. This work was funded in part by the USDA National Institute of Food and Agriculture, grant number 2018-67015-28199, and the National Science Foundation (NSF), grant IOS-1744309, and NSF PPoSS award number 2118709 (E.G. and P.P.). This work was funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). G.Bourque is supported by a Canada Research Chair Tier 1 award, a FRQ-S, Distinguished Research Scholar award and by the World Premier International Research Center Initiative (WPI), MEXT, Japan. J.S. was supported by the Carlsberg Foundation. This work was funded in part by intramural funding at the National Institute of Standards and Technology. E.E.E., D.H. and E.D.J. are investigators of the Howard Hughes Medical Institute. This work was funded in part by an Oxford Nanopore Research grant (SC20130149) awarded to M. Akeson, University of California Santa Cruz. This work was funded in part by Wellcome Trust award numbers WT104947/Z/14/Z, WT222155/Z/20/Z and WT108749/Z/15/Z. This work was funded in part by a Juan de la Cierva fellowship grant (IJC2020-045916-I) funded by MCIN/AEI/ 10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work was funded in part by the Novo Nordisk Foundation (NNF21OC0069089). S.H. acknowledges funding from the Central Innovation Programme (ZIM) for SMEs of the Federal Ministry for Economic Affairs and Energy of Germany. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D and 031A538A). This work was funded in part by the German Federal Ministry of Education and Research (BMBF) (031L0184A) and the European Commission, Innovative training network (ITN) (956229). W.-W.L. was supported in part by the Government Scholarship to Study Abroad (GSSA) from the Ministry of Education of Taiwan.

## Author information

Author notes

1. These authors contributed equally: Wen-Wei Liao, Mobin Asri, Jana Ebler

## Authors and Affiliations

1. Department of Genetics, Yale University School of Medicine, New Haven, CT, USA

   Wen-Wei Liao, Shuangjia Lu & Ira M. Hall

2. Center for Genomic Health, Yale University School of Medicine, New Haven, CT, USA

   Wen-Wei Liao, Shuangjia Lu & Ira M. Hall

3. Division of Biology and Biomedical Sciences, Washington University School of Medicine, St. Louis, MO, USA

   Wen-Wei Liao

4. Genomics Institute, University of California, Santa Cruz, CA, USA

   Mobin Asri, Marina Haukness, Glenn Hickey, Julian K. Lucas, Jean Monlong, Xian H. Chang, Jordan M. Eizenga, Charles Markello, Adam M. Novak, Hugh E. Olsen, Trevor Pesout, Jouni Sirén, Mark Diekhans, David Haussler, Karen H. Miga & Benedict Paten

5. Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany

   Jana Ebler, Daniel Doerr, Peter Ebert, Hugo Magalhães, Pierre Marijon & Tobias Marschall

6. Center for Digital Medicine, Heinrich Heine University, Düsseldorf, Germany

Jana Ebler, Daniel Doerr, Peter Ebert, Hugo Magalhães, Pierre Marijon & Tobias Marschall

7. Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO, USA

Haley J. Abel

8. Institute of Genetics and Biophysics, National Research Council, Naples, Italy

Silvia Buonaiuto & Vincenza Colonna

9. Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

Haoyu Cheng, Justin Chu, Xiaowen Feng & Heng Li

10. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Haoyu Cheng, Xiaowen Feng & Heng Li

11. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

Vincenza Colonna, Christian Fischer, Andrea Guarracino, Pjotr Prins, Flavia Villani & Erik Garrison

12. McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA

Robert S. Fulton, Chad Tomlinson, Lucinda L. Antonacci-Fulton, Sarah Cody, Allison A. Regier & Ting Wang

13. Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA

Robert S. Fulton & Ting Wang

14. Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Copenhagen, Denmark

Shilpa Garg

15. Quantitative Life Sciences, McGill University, Montréal, Québec, Canada

Cristian Groza

16. Genomics Research Centre, Human Technopole, Milan, Italy

Andrea Guarracino

17. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

William T. Harvey, Katherine M. Munson, David Porubsky, Mitchell R. Vollger, Carl A. Baker, Kendra Hoekzema, Jennifer Kordosky, Alexandra P. Lewis & Evan E. Eichler

18. Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany

Simon Heumos

19. Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen, Germany

Simon Heumos

20. Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridge, UK

Kerstin Howe & Jonathan M. D. Wood

21. Northeastern University, Boston, MA, USA

Miten Jain

22. Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA

Tsung-Yu Lu & Mark J. P. Chaisson

23. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

Fergal J. Martin, Konstantinos Billis, Susan Fairley, Adam Frankish, Carlos Garcia Giron, Leanne Haggerty, Thibaut Hourlier, Jan O. Korbel, Francesca Floriana Tricomi & Paul Flicek

24. Coriell Institute for Medical Research, Camden, NJ, USA

Matthew W. Mitchell

25. Department of Computer Science, University of Pisa, Pisa, Italy

Moses Njagi Mwaniki

26. Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark

Jonas A. Sibbesen

27. Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA, USA

Mitchell R. Vollger

28. Google, Mountain View, CA, USA

Gunjan Baid, Anastasiya Belyaeva, Andrew Carroll, Pi-Chuan Chang, Daniel E. Cook, Alexey Kolesnikov, Maria Nattestad & Kishwar Shafin

29. Barrett and O’Connor Washington Center, Arizona State University, Washington, DC, USA

Robert M. Cook-Deegan

30. Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, CA, USA

Omar E. Cornejo & Samuel Sacco

31. Core Unit Bioinformatics, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany

Peter Ebert

32. Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA

Olivier Fedrigo, Giulio Formenti, Jacquelyn Mountcastle & Erich D. Jarvis

33. National Institutes of Health (NIH)–National Human Genome Research Institute, Bethesda, MD, USA

Adam L. Felsenfeld, Baergen I. Schultz, Michael W. Smith & Heidi J. Sofia

34. Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

Yan Gao

35. Institute for Society and Genetics, College of Letters and Science, University of California, Los Angeles, CA, USA

Nanibaa' A. Garrison

36. Institute for Precision Health, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

Nanibaa' A. Garrison

37. Division of General Internal Medicine and Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

Nanibaa' A. Garrison

38. Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA

Richard E. Green

39. Dovetail Genomics, Scotts Valley, CA, USA

Richard E. Green

40. Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

Hanlee P. Ji & HoJoon Lee

41. Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Eimear E. Kenny

42. Program in Bioethics and Institute for Human Genetics, University of California, San Francisco, CA, USA

Barbara A. Koenig

43. Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

Jan O. Korbel

44. Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Sergey Koren, Ann McCartney, Sergey Nurk, Mikko Rautiainen, Arang Rhie, Brian Walenz & Adam M. Phillippy

45. Computer Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

Santiago Marco-Sola

46. Departament d'Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Barcelona, Spain

Santiago Marco-Sola

47. Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

Jennifer McDaniel, Nathan D. Olson, Justin Wagner & Justin M. Zook

48. Department of Public Health Sciences, University of California, Davis, CA, USA

Alice B. Popejoy

49. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

Daniela Puiu & Aleksey V. Zimin

50. Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany

Ashley D. Sanders

51. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Valerie A. Schneider & Françoise Thibaud-Nissen

52. Al Jalila Genomics Center of Excellence, Al Jalila Children's Specialty Hospital, Dubai, UAE

Ahmad N. Abou Tayoun

53. Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, UAE

Ahmad N. Abou Tayoun

54. Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA

Aleksey V. Zimin

55. Department of Human Genetics, McGill University, Montréal, Québec, Canada

Guillaume Bourque

56. Canadian Center for Computational Genomics, McGill University, Montréal, Québec, Canada

Guillaume Bourque

57. Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan

Guillaume Bourque

58. Howard Hughes Medical Institute, Chevy Chase, MD, USA

Evan E. Eichler, David Haussler & Erich D. Jarvis

59. Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA

Erich D. Jarvis

## Contributions

Pangenome empirical analysis and pangenome quality control: W.-W.L., D.D., M.H., G.H., C.M., J. Monlong, H.J.A., J.M.Z., E.E.E., T.M., I.M.H., P.M. and J.W. Paper writing: W.-W.L., M.A., J.E., D.D., M.H., G.H., S.L., J. Monlong, R.S.F., S.G., T.-Y.L., M.W.M., A.M.N., H.E.O., T.P., J.A.S., M.R.V., G. Bourque, K.H.M., E.G., T.M., I.M.H., B.P., R.E.G. and L.H. Paper editing: W.-W.L., D.D., M.H., G.H., X.H.C., H.C., A.G., A.M.N., P.P., A.M.P., E.E.E., E.D.J., K.H.M., E.G., E.E.K., T.M., I.M.H., H.L., B.P., O.E.C., P.E., G.F., A.N.A.T. and A.V.Z. Assembly creation: M.A., J.K.L., H.C., A.M.P., H.L., D. Puiu, A.A.R. and A.V.Z. Assembly quality control and assembly reliability analysis: M.A., J.K.L., H.C., J.C., S.G., K. Howe, T.P., D. Porubsky, C.T., M.R.V., A.M.P., J.M.Z., E.E.E., K.H.M., H.L., R.E.G., S.K., J. McDaniel, S.N., N.D.O., D. Puiu, M.R., A.A.R., A.R., V.A.S., K.S., F.T.-N., J.W., B.W., J.M. D.W. and A.B.P. Pangenome applications (structural variants): J.E., G.H., H.J.A., W.T.H., P.P., E.E.E., T.M., H.P.J. and H.M. Pangenome graph creation: D.D., G.H., A.G., S.H., M.N.M., F.V., E.G., Y.G. and S.M.-S. Data coordination and management: M.H., J.K.L., R.S.F., W.T.H., M.J., C.T., A.M.P., E.D.J., K.H.M., T.W., L.L.A.-F., S.C., M.D., S.F. and R.E.G. Transcriptome and annotation: M.H., J.M.E., F.J.M., M.R.V., M.J.P.C., K.B., M.D., A.F., C.G.G., L.H., T.H. and F.F.T. Pangenome applications (small variants): G.H., J. Monlong, C.M., A.M.N., P.P., J.M.Z., G. Baid, A.B., A.C., P.-C.C., D.E.C., H.P.J., A.K., M.N., K.S. and J.W. Pangenome visualization and complex loci analysis: S.L., J.C., C.F. and A.G. Population genetic analysis: S.B., A.G. and V.C. Sample selection: X.F., K.M.M., A.M.P., E.E.K., E.E.E., K.H.M., S.F. and J.O.K. Sequencing: R.S.F., M.J., M.W.M., K.M.M., H.E.O., A.M.P., E.E.E., E.D.J., K.H.M., C.A.B., O.F., R.E.G., K. Hoekzema, J.O.K., J.K., A.P.L., J. Mountcastle, S.S. and A.D.S. Pangenome applications (ChIP-seq analysis): C.G. and G. Bourque. Principal investigator and laboratory organizer within the HPRC: M.J., M.W.M., M.J.P.C., P.F., E.E.K., A.M.P., E.E.E., D.H., E.D.J., K.H.M., T.W., T.M., B.P., R.E.G. and V.A.S. Pangenome applications (VNTR analysis): T.-Y.L. and M.J.P.C. Pangenome applications (RNA-seq analysis): J.A.S. and J.M.E. Development of algorithms and software: J.S., H.P.J., H.L., G.H., A.M.N., P.P., A.G., S.H., D.D., M.N.M., F.V., E.G., Y.G., S.M.-S., H.L., J.M.E., J.A.S., B.P., T.M. and X.H.C. Ethical, legal and social implications: R.M.C.-D., N.A.G.,

B.A.K., A.M. and A.B.P. Programme organization: A.L.F., B.I.S., M.W.S. and H.J.S.


### Corresponding authors

Correspondence to Erik Garrison, Tobias Marschall, Ira M. Hall, Heng Li or Benedict Paten.


## Ethics declarations

### Competing interests

E.E.E. is a scientific advisory board (SAB) member of Variant Bio. P.F is a member of the SABs of Fabric Genomics and Eagle Genomics. E.E.K. is a member of the SAB of Encompass Biosciences, Foresite Labs and Galateo Bio and has received personal fees from Regeneron Pharmaceuticals, 23&Me and Illumina. A.B., A.C., P.-C.C., D.E.C., G.Baid, A.K., M.N. and K.S. are employees of Google and own Alphabet stock as part of the standard compensation package.

## Peer review

### Peer review information

*Nature* thanks Melissa Gymrek, Martin Mascher and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

## Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Extended data figures and tables

## Extended Data Fig. 1 Characterizing uncovered reference bases using peri/centromeric annotation and evaluating the completeness of different satellite families.

We characterized the regions not covered by the assembly alignments to the T2T-CHM13 (v.2.0) reference and also investigated the completeness of the peri/centromeric satellites across all HPRC assemblies. We characterized these regions using the peri/centromeric annotation available for the T2T-CHM13 (v.2.0) reference. We made separate bar plots for male and female samples to exclude chromosome X for the paternal assemblies of male samples and exclude chromosome Y for all other assemblies. Panels **a** and **b** indicate that on average ~90% of the uncovered bases are located in peri/centromeric regions with the active/inactive alpha satellites and human satellite 3 comprising ~50% of these bases, mainly due to their highly repetitive composition and also higher frequency compared to other satellites. Other centromeric satellites, centromeric transition regions, and rDNA arrays accounted for another ~40% of the uncovered bases on average. Panels **c** and **d** display the average lengths of uncovered regions located within each satellite family. Panels **e** and **f** show what percentage of each satellite family was covered by at least one assembly alignment. The most complete centromeric regions (~90% coverage) are divergent/monomeric alpha satellites, gamma satellites and centromeric transition regions. The rDNA arrays have been covered by ~8% on average, which made them the least completely assembled repeat arrays.

## Extended Data Fig. 2 Segmental duplication reliability.

**a**, Average number of Mbp per haplotype of correctly or incorrectly assembled SDs lifted from T2T-CHM13 (v.2.0). **b**, The features of the most identical and longest overlapping SDs for each type of assembly error calculated in 5 kbp windows.

## Extended Data Fig. 3 The differences in pangenome graph construction methods for Minigraph, MC, and PGGB.

**a**, Two haplotypes ($H_1$ and $H_2$) vary in copy number of a chromosomal segment S. The $S_1$, $S_2$, and $S_3$ segments are highly similar with only a SNP or a small indel. **b**, Pangenome graph structures for Minigraph, MC, and PGGB. Minigraph used $H_1$ as an initial backbone and then augmented with SVs (≥50 bp) from $H_2$, such that the SNP in $S_2$ is not represented in the pangenome graph. MC added small variants (<50 bp) to the pangenome graph constructed by Minigraph. PGGB used a symmetric, all-by-all alignment of haplotypes to build a pangenome graph whose structure is not affected by the order of inputs (unlike Minigraph and MC). The critical difference in graph construction is that, due to ambiguous pairwise relationships of paralogs, PGGB tends to collapse copy-number polymorphic loci like segmental duplications and VNTRs into a single copy through which haplotypes loop, while Minigraph and MC do not.

## Extended Data Fig. 4 HiFi read depth of on- and off-target edges in the MC graph.

Left: fraction of reads aligned to the pangenome graph after filtering low-quality alignments. Middle: read depth distribution of on-target edges. Right: read depth distribution of off-target edges. Samples are sorted by sequencing coverage (Supplementary Table 1).

## Extended Data Fig. 5 Gene mapping in the pangenome graphs.

The first three show the percentage of protein-coding genes from GENCODE (v.38) able to be mapped in the gene annotation sets from Ensembl, CAT run on the MC graph based on GRCh38, and CAT run on the PGGB graph. The second three show the percentage of noncoding genes from GENCODE (v.38) able to be mapped on the same annotation sets.

## Extended Data Fig. 6 Structural haplotypes of *CYP2D6* and *CYP2D7* from the MC graph.

**a**, Locations of *CYP2D6* and *CYP2D7* within the graph. The colour gradient is based on the precise relative position of each gene; green, head of a gene; blue, end of a gene. **b**, Different structural haplotypes take different paths

through the graph. The colour gradient and lines show the path of each allele; red, start of a path; blue, end of a path. **c**, Frequency and linear structural visualization of all structural haplotypes called by the graph among 90 haploid assemblies.

## Extended Data Fig. 7 Performance comparison of pangenome-based variant calling and read mapping across populations.

**a**, Number of variants with at least one alternate allele (i.e. excluding homozygous for the reference allele) for each in the 1KG samples. The number of variants in the 1KG callset (x-axis) are compared to the variants found when aligning reads to the HPRC pangenome and calling variants with DeepVariant (y-axis). Points (samples) are coloured by their super-population label from the 1KG. **b**, The proportion of mapped reads that align perfectly (y-axis) is shown for a subset of samples from the 1KG, ordered by the number of variants called (x-axis). Two mapping approaches are compared: mapping short reads to GRCh38 with BWA (green); mapping to the HPRC pangenome with Giraffe (orange). The samples were selected to span the x-axis.

## Extended Data Fig. 8 Improved genotyping in the challenging medically-relevant gene *RHCE*.

**a**, Gene annotation of part of the *RHCE* gene. **b**, Genotyping performance in this region for three approaches (horizontal panels). The top panel, using the HPRC pangenome, shows the best performance with most variants being true positives (TP, blue points) based on the CMRG (v.1.0) truth set while more other methods have a higher number of false negatives (FN, red points). **c**, Allele frequency across 2,504 unrelated individuals of the 1KG. The HPRC-Giraffe-DeepVariant calls show higher frequencies. In particular, the gene-converted alleles, at about 25.406-25.410 Mbp, are observed at ~25% frequency, similar to estimates from the HPRC haplotypes (Fig. 5a–c). **d**,**e**, A pangenomic view of the gene-converted region showing 1 of 4 haplotypes in the HPRC pangenome supporting the non-reference alleles. The inclusion of this haplotype in the HPRC

pangenome enables short sequencing reads, here from HG002, to map along this gene-converted haplotype.

## Extended Data Fig. 9 Leave-one-out experiment.

A leave-one-out experiment was conducted by repeatedly removing one of the assembly-samples from the panel VCF and genotyping it based on the remaining samples. Plots show the resulting weighted genotype concordances for different variant allele classes. **a**, weighted genotype concordances are stratified by graph complexity: biallelic regions of the MC graph include only bubbles with two branches, and multiallelic regions include all bubbles with > 2 different alternative paths defined by the 88 haplotypes. **b**, results of the same experiment stratified by different genomic regions defined by the GIAB.

## Extended Data Fig. 10 Additional applications supported by the pangenome reference.

**a**, Performance of read alignment in VNTR regions using the MC graph versus GRCh38. All statistics are expressed relative to the total number of reads simulated from each genome. **b**, Performance of RNA-seq read alignment. Mapping rate and false discovery rate are stratified by mapping quality producing the curves shown. The MC graph is compared to a graph derived from the 1KG variant calls and to GRCh38. Each reference is augmented with splice junctions. vg mpmap was used to map to the graphs, and STAR was used to map to the linear reference. **c**, Proportion of all ChIP-seq peaks that are called only in the MC graph. Each data point represents samples that were assayed for H3K4me1, H3K27ac histone marks or chromatin accessibility using ATAC-seq. **d**, H3K4me1 peaks that overlap an SV for which the sample is heterozygous. The reads within the peak are partitioned between the SV or reference allele. The red boundary represents regions where a binomial test assigns a peak to the SV allele, both alleles, or the reference allele.

## Extended Data Fig. 11 Number of SVs per sample in the HPRC PanGenie filtered set as well as the 1KG Illumina calls for all

### 3,202 1KG samples.

Samples are coloured by superpopulation. The left plot excludes the african superpopulation, while the right plot shows the same results including african samples and including the assembly samples present in the graph (marked by a black circle).

## Supplementary information

### Supplementary Figures

Supplementary Figs. 1–49 and their legends.

### Reporting Summary

### Peer Review File

### Supplementary Tables

Supplementary Tables 1–22 and their legends.

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

# About this article

## Cite this article

Liao, WW., Asri, M., Ebler, J. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023). https://doi.org/10.1038/s41586-023-05896-x

- Received: 09 July 2022
- Accepted: 28 February 2023
- Published: 10 May 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05896-x

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

## Human pangenome supports analysis of complex genomic regions

- Arya Massarat
- Melissa Gymrek
- Hákon Jónsson

Nature News & Views Forum 10 May 2023

## Increased mutation and gene conversion within human segmental duplications

- Mitchell R. Vollger
- Philip C. Dishuck
- Evan E. Eichler

Nature Article Open Access 10 May 2023

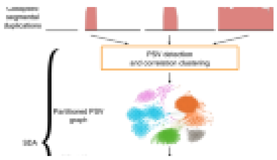
## ‘Pangenome’ aims to capture the breadth of human diversity

- Nick Petrić Howe
- Shamini Bundell

Nature Nature Podcast 10 May 2023

---

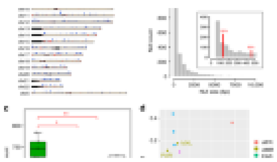
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05896-x



## Long-read sequence and assembly of segmental duplications

17 December 2018

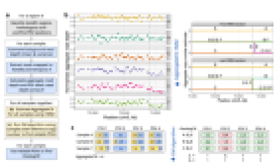
Mitchell R. Vollger, Philip C. Dishuck, … Evan E. Eichler



## De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations

02 August 2018

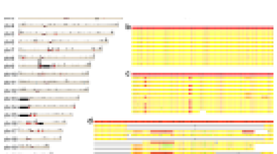
Karen H. Y. Wong, Michal Levy-Sakin & Pui-Yan Kwok



## Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing

09 June 2022

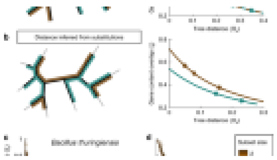
Timofey Prodanov & Vikas Bansal



## [Genome maps across 26 human populations reveal population-specific patterns of structural variation]

04 March 2019

Michal Levy-Sakin, Steven Pastor, … Pui-Yan Kwok



## [Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence]

26 November 2019

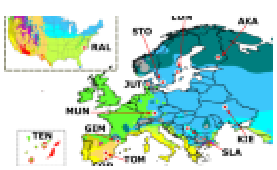
Jaime Iranzo, Yuri I. Wolf, … Itamar Sela



## [Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in Drosophila and humans]

30 December 2021

Dan Zhang, Liang Leng, … Yong E. Zhang

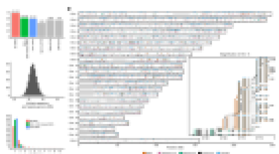


## [Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive

[signatures in Drosophila]

12 April 2022

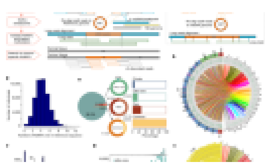
Gabriel E. Rech, Santiago Radío, … Josefa González



[The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes]

22 April 2020

Laurits Skov, Moisès Coll Macià, … Kari Stefansson



[Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes]

05 October 2022

Wei Wei, Katherine R. Schon, … Patrick F. Chinnery

- Article
- Open Access
- Published: 10 May 2023

# Increased mutation and gene conversion within human segmental

# duplications


- Mitchell R. Vollger,
- Philip C. Dishuck,
- William T. Harvey,
- William S. DeWitt,
- Xavi Guitart,
- Michael E. Goldberg,
- Allison N. Rozanski,
- Julian Lucas,
- Mobin Asri,
- Human Pangenome Reference Consortium,
- Katherine M. Munson,
- Alexandra P. Lewis,
- Kendra Hoekzema,
- Glennis A. Logsdon,
- David Porubsky,
- Benedict Paten,
- Kelley Harris,
- PingHsun Hsieh &
- …
- Evan E. Eichler

Show authors



*Nature* **volume 617**, pages 325–334 (2023)



- 332 Altmetric

- Metrics details


## Abstract


Single-nucleotide variants (SNVs) in segmental duplications (SDs) have not been systematically assessed because of the limitations of mapping short-read sequencing data[1,2]. Here we constructed 1:1 unambiguous alignments

spanning high-identity SDs across 102 human haplotypes and compared the pattern of SNVs between unique and duplicated regions[3,4]. We find that human SNVs are elevated 60% in SDs compared to unique regions and estimate that at least 23% of this increase is due to interlocus gene conversion (IGC) with up to 4.3 megabase pairs of SD sequence converted on average per human haplotype. We develop a genome-wide map of IGC donors and acceptors, including 498 acceptor and 454 donor hotspots affecting the exons of about 800 protein-coding genes. These include 171 genes that have 'relocated' on average 1.61 megabase pairs in a subset of human haplotypes. Using a coalescent framework, we show that SD regions are slightly evolutionarily older when compared to unique sequences, probably owing to IGC. SNVs in SDs, however, show a distinct mutational spectrum: a 27.1% increase in transversions that convert cytosine to guanine or the reverse across all triplet contexts and a 7.6% reduction in the frequency of CpG-associated mutations when compared to unique DNA. We reason that these distinct mutational properties help to maintain an overall higher GC content of SD DNA compared to that of unique DNA, probably driven by GC-biased conversion between paralogous sequences[5,6].

# Main

The landscape of human SNVs has been well characterized for more than a decade in large part owing to wide-reaching efforts such as the International HapMap Project and the 1000 Genomes Project[7,8]. Although these consortia helped to establish the genome-wide pattern of SNVs (as low as 0.1% allele frequency) and linkage disequilibrium on the basis of sequencing and genotyping thousands of human genomes, not all parts of the human genome could be equally ascertained. Approximately 10–15% of the human genome[8] has remained inaccessible to these types of analysis either because of gaps in the human genome sequence or, more frequently, the low mapping quality associated with aligning short-read whole-genome sequencing data. This is because short-read sequence data are of insufficient length (<200 base pairs (bp)) to unambiguously assign reads and, therefore, variants to specific loci[9]. Although certain classes of large, highly identical repeats (for example, α-satellites in centromeres) were readily recognized, others, especially SDs[1]

and their 859 associated genes[10], in euchromatin were much more problematic to recognize.

Operationally, SDs are defined as interchromosomal or intrachromosomal homologous regions in any genome that are >1 kbp in length and >90% identical in sequence[1,11]. As such regions arise by duplication as opposed to retrotransposition, they were initially difficult to identify and early versions of the human genome sequence had either missed or misassembled these regions owing to their high sequence identity[12,13]. Large-insert BAC clones ultimately led to many of these regions being resolved. Subsequent analyses showed that SDs contribute disproportionately to copy number polymorphisms and disease structural variation[9,14], are hotspots for gene conversion[15], are substantially enriched in GC-rich DNA and Alu repeats[16,17], and are transcriptionally diverse leading to the emergence, in some cases, of human-specific genes thought to be important for human adaptation[18,19,20,21]. Despite their importance, the pattern of SNVs among humans has remained poorly characterized. Early on, paralogous sequence variants were misclassified as SNVs[2] and, as a result, later high-identity SDs became blacklisted from SNV analyses because short-read sequence data could not be uniquely placed[22,23]. This exclusion has translated into a fundamental lack of understanding in mutational processes precisely in regions predicted to be more mutable owing to the action of IGC[24,25,26,27,28]. Previously, we noted an increase in SNV density in duplicated regions when compared to unique regions of the genome on the basis of our comparison of GRCh38 and the complete telomere-to-telomere (T2T) human reference genome[10]. Leveraging high-quality phased genome assemblies from 47 humans generated as part of the Human Pangenome Reference Consortium (HPRC)[3], we sought to investigate this difference more systematically and compare the SNV landscape of duplicated and unique DNA in the human genome revealing distinct mutational properties.

## Strategy and quality control

Unlike previous SNV discovery efforts, which catalogued SNVs on the basis of the alignment of sequence reads, our strategy was assembly driven (Extended Data Fig. 1). We focused on the comparison of 102 haplotype-

resolved genomes (Supplementary Table 1) generated as part of the HPRC ($n = 94$) or other efforts ($n = 8$)[3,4,12,29] in which phased genome assemblies had been assembled using high-fidelity (HiFi) long-read sequencing[30]. The extraordinary assembly contiguity of these haplotypes (contig N50, defined as the sequence length of the shortest contig at 50% of the total assembly length, > 40 Mbp) provided an unprecedented opportunity to align large swathes (>1 Mbp) of the genome, including high-identity SD repeats anchored by megabases of synteny.

As SD regions are often enriched in assembly errors even among long-read assemblies[3,4,31], we carried out a series of analyses to assess the integrity and quality of these regions in each assembled haplotype. First, we searched for regions of collapse[11] by identifying unusual increases or decreases in sequence read depth[3]. We determine that, on average, only 1.64 Mbp (1.37%) of the analysed SD sequence was suspect owing to unusually high or low sequence read depth on the basis of mapping of underlying read data — as such patterns are often indicative of a misassembly[3] (Methods). Next, for all SD regions used in our analysis we compared the predicted copy number by Illumina sequence read depth with the sum based on the total copy number from the two assembled haplotypes. These orthogonal copy number estimates were highly correlated (Pearson's $R = 0.99$, $P < 2.2 \times 10^{-16}$; Supplementary Fig. 1) implying that most SD sequences in the assemblies have the correct copy number. To confirm these results in even the most difficult to assemble duplications, we selected 19 of the largest and most identical SDs across 47 haplotypes for a total of 893 tests. These estimates were also highly correlated (Pearson's $R = 0.99$, $P < 2.2 \times 10^{-16}$; Supplementary Figs. 2 and 3), and of the 893 tests conducted, 756 were identical. For the 137 tests for which estimates differed, most ($n = 125$) differed by only one copy. Finally, most of these discrepancies came from just three large (>140 kbp) and highly identical (>99.3%) SDs (Supplementary Fig. 3).

To validate the base-level accuracy, we next compared the quality value for both SD and unique sequences using Illumina sequencing data for 45 of the HPRC samples (Methods). Both unique (average quality value = 59 s.d. 1.9) and SD (average quality value = 53 s.d. 1.9) regions are remarkably high quality, which in the case of SDs translates into less than 1 SNV error every

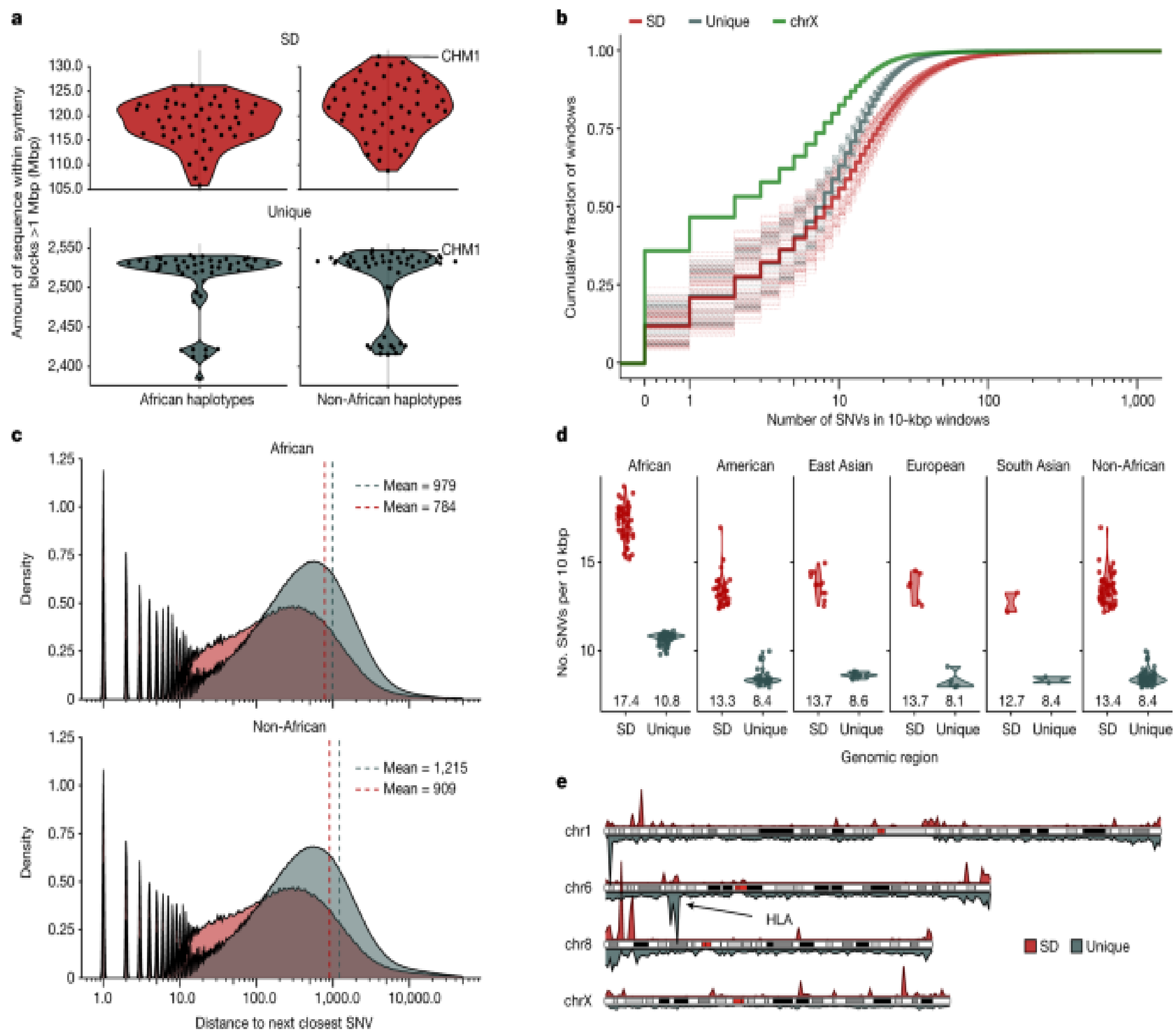
200 kbp (Supplementary Fig. 4). We further show that these high-quality assembles result in accurate variant calls (Supplementary Notes and Supplementary Figs. 5–9). We also assessed the contiguity of the underlying assemblies using a recently developed tool, GAVISUNK, which compares unique *k*-mer distributions between HiFi-based assemblies and orthogonal Oxford Nanopore Technologies sequencing data from the same samples. We found that, on average, only 0.11% of assayable SD sequence was in error compared to 0.14% of unique regions assayed (Supplementary Table 2), implying high and comparable assembly contiguity. As a final control for potential haplotype-phasing errors introduced by trio HiFi assembly of diploid samples, we generated deep Oxford Nanopore Technologies and HiFi data from a second complete hydatidiform mole (CHM1) for which a single paternal haplotype was present and applied a different assembly algorithm[32] (Verkko 1.0; Extended Data Fig. 2). We show across our many analyses that the results from the CHM1 Verkko assembly are consistent with individual haplotypes obtained from diploid HPRC samples produced by trio hifiasm[3,32] (Supplementary Fig. 10). We therefore conclude that phasing errors have, at most, a negligible effect on our results and that most (>98%) SDs analysed were accurately assembled from multiple human genomes allowing the pattern of SNV diversity in SDs to be systematically interrogated.

# Increased SNV density in SD regions

To assess SNVs, we limited our analysis to portions of the genome where a 1:1 orthologous relationship could be unambiguously assigned (as opposed to regions with extensive copy number variation). Using the T2T-CHM13 reference genome, we aligned the HPRC haplotypes requiring alignments to be a minimum of 1 Mbp in length and carry no structural variation events greater than 10 kbp (Methods and Extended Data Fig. 1). Although the proportion of haplotypes compared for any locus varied (Fig. 1a), the procedure allowed us to establish, on average, 120.2 Mbp 1:1 fully aligned sequence per genome for SD regions out of a total of 217 Mbp from the finished human genome (T2T-CHM13 v1.1). We repeated the analysis for ‘unique’ (or single-copy) regions of the genome and recovered by comparison 2,508 Mbp as 1:1 alignments (Fig. 1a). All downstream analyses were then carried out using this orthologous alignment set. We first

compared the SNV diversity between unique and duplicated regions excluding suboptimal alignments mapping to tandem repeats or homopolymer stretches. Overall, we observe a significant 60% increase in SNVs in SD regions (Methods; Pearson's chi-squared test with Yates's continuity correction $P < 2.2 \times 10^{-16}$; Fig. 1b). Specifically, we observe an average of 15.3 SNVs per 10 kbp versus 9.57 SNVs per 10 kbp for unique sequences (Fig. 1d). An empirical cumulative distribution comparing the number of SNVs in 10-kbp windows between SD and unique sequence confirms that this is a general property and not driven simply by outliers. The empirical cumulative distribution shows that more than half of the SD sequences have more SNVs than their unique counterparts (Fig. 1b). Moreover, for all haplotypes we divided the unique portions of the genome into 125-Mbp bins and found that all SD bins of equivalent size have more SNVs than any of the bins of unique sequence (empirical $P$ value < 0.0005; Extended Data Fig. 3). This elevation in SNVs is only modestly affected by the sequence identity of the underlying SDs (Pearson's correlation of only 0.008; Supplementary Fig. 11). The increase in SNVs (60%) in SDs is greater than that in all other assayable classes of repeats: Alu (23%), L1 (−9.4%), human endogenous retroviruses (−9.4%) and ancient SDs for which the divergence is greater than 10% (12%) (Extended Data Fig. 4 and Supplementary Table 3). We find, however, that SNV density correlates with increasing GC content (Supplementary Fig. 12) consistent with Alu repeats representing the only other class of common repeat to show an elevation.

**Fig. 1: Increased single-nucleotide variation in SDs.**




**a**, The portion of the human genome analysed for SD (red) and unique (blue) regions among African and non-African genomes. Shown are the number of megabase pairs aligned in 1:1 syntenic blocks to T2T-CHM13 v1.1 for each assembled haplotype. Data are shown as both a single point per haplotype originating from a single individual and a smoothed violin plot to represent the population distribution. **b**, Empirical cumulative distribution showing the number of SNVs in 10-kbp windows in the syntenic regions stratified by unique (grey), SD (red) and the X chromosome (chrX; green). Dashed lines represent individual haplotypes and thick lines represent the average trend of all the data. **c**, Distribution of the average distance to the next closest SNV in SD (red) and unique (grey) space separating African (top) and non-African (bottom) samples. Dashed vertical lines are drawn at the mean of each distribution. **d**, Average number of SNVs per 10-kbp window in SD (red) versus unique (grey) space by superpopulation and with mean value shown

underneath each violin. The non-African column represents an aggregation of the data from all non-African populations in this study. **e**, Density of SNVs in 10 bp of each other for SD (top, red) and unique (bottom, grey) regions for chromosomes 1, 6, 8 and X comparing the relative density of known (for example, HLA) and new hotspots of single-nucleotide variation.

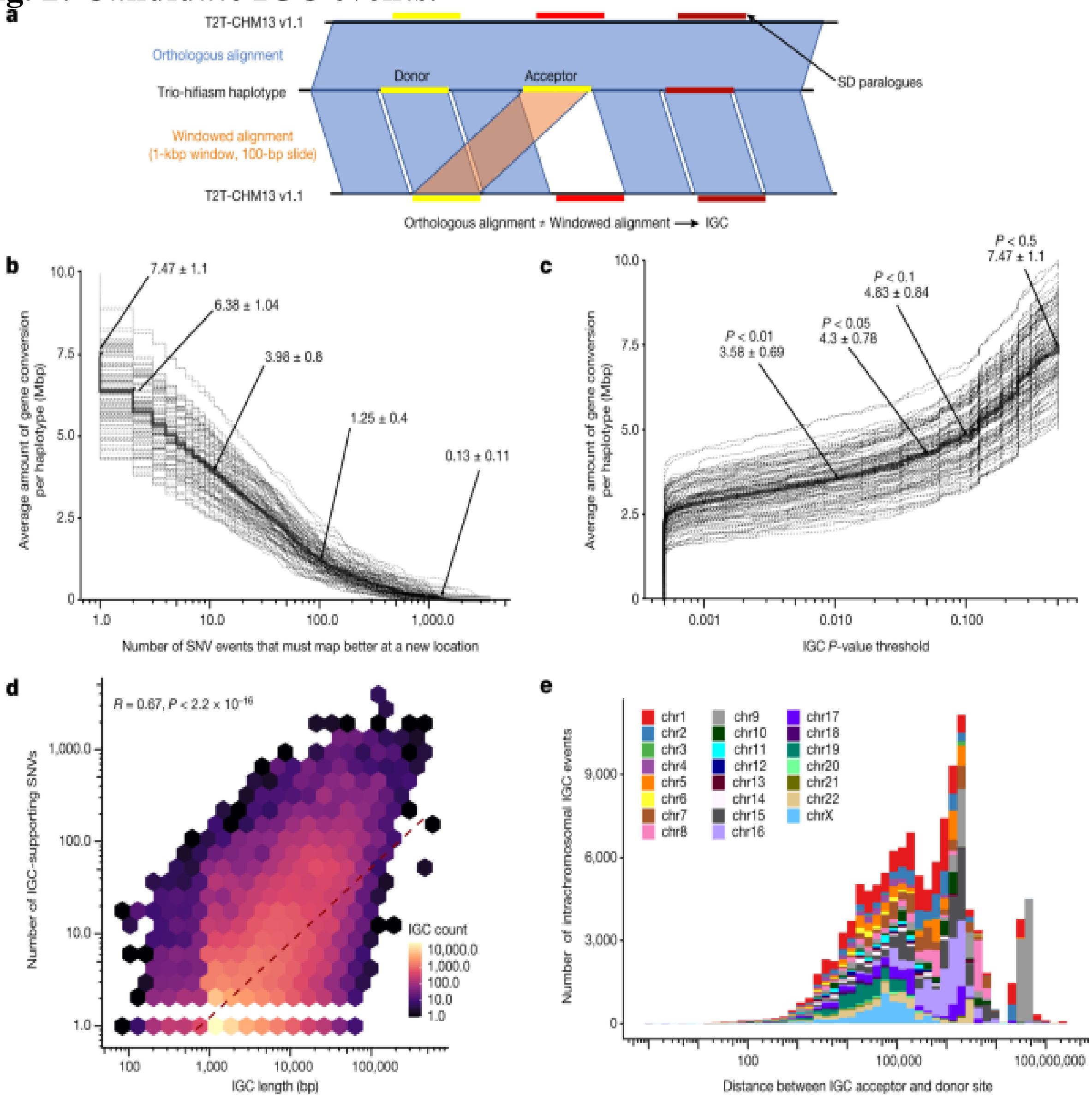
Previous publications have shown that African haplotypes are genetically more diverse, having on average about 20% more variant sites compared to non-African haplotypes[8]. To confirm this observation in our data, we examined the number of SNVs per 10 kbp of unique sequence in African versus non-African haplotypes (Fig. 1c,d) and observed a 27% (10.8 versus 8.5) excess in African haplotypes. As a result, among African haplotypes, we see that the average distance between SNVs (979 bp) is 19.4% closer than in non-African haplotypes (1,215 bp), as expected[8,12]. African genomes also show increased variation in SDs, but it is less pronounced with an average distance of 784 bases between consecutive SNVs as compared to 909 bases in non-African haplotypes (13.8%). Although elevated in African haplotypes, SNV density is higher in SD sequence across populations and these properties are not driven by a few sites but, once again, are a genome-wide feature. We put forward three possible hypotheses to account for this increase although note these are not mutually exclusive: SDs have unique mutational mechanisms that increase SNVs; SDs have a deeper average coalescence than unique parts of the genome; and differences in sequence composition (for example, GC richness) make SDs more prone to particular classes of mutation.

# Putative IGC

One possible explanation for increased diversity in SDs is IGC in which sequence that is orthologous by position no longer shares an evolutionary history because a paralogue from a different location has ‘donated’ its sequence through ectopic template-driven conversion[33], also known as nonallelic gene conversion[27]. To identify regions of IGC, we developed a method that compares two independent alignment strategies to pinpoint regions where the orthologous alignment of an SD sequence is inferior to an independent alignment of the sequence without flanking information (Fig. 2a and Methods). We note several limitations of our approach (Supplementary

Notes); however, we show that our high-confidence IGC calls (20+ supporting SNVs) have strong overlap with other methods for identifying IGC (Supplementary Notes and Supplementary Fig. 13). Using this approach, we created a genome-wide map of putative large IGC events for all of the HPRC haplotypes for which 1:1 orthologous relationships could be established (Fig. 2).

**Fig. 2: Candidate IGC events.**




**a**, Method to detect IGC. The assembled human haplotype query sequence from 1:1 syntenic alignments was fragmented into 1-kbp windows in 100-bp increments and realigned back to T2T-CHM13 v1.1 independent of the

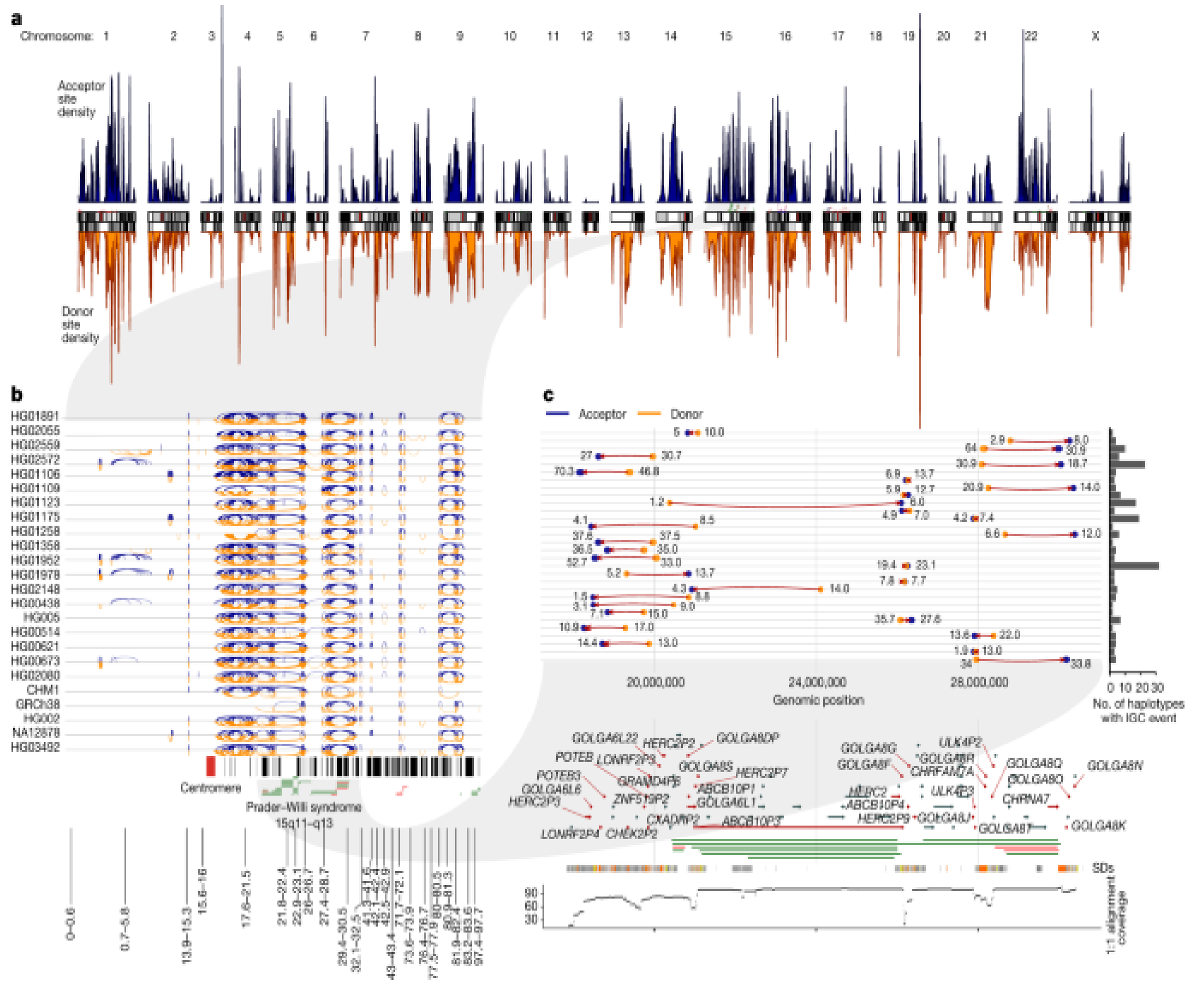
flanking sequence information using minimap2 v2.24 to identify each window's single best alignment position. These alignments were compared to their original syntenic alignment positions, and if they were not overlapping, we considered them to be candidate IGC windows. Candidate IGC windows were then merged into larger intervals and realigned when windows were overlapping in both the donor and the acceptor sequence. We then used the CIGAR string to identify the number of matching and mismatching bases at the 'donor' site and compared that to the number of matching and mismatching bases at the acceptor site determined by the syntenic alignment to calculate the number of supporting SNVs. **b**, The amount of SDs (in megabase pairs) predicted to be affected by IGC per haplotype, as a function of the minimum number of SNVs that support the IGC call. Dashed lines represent individual haplotypes and the solid line represents the average. **c**, Empirical cumulative distribution of the megabase pairs of candidate IGC observed in HPRC haplotypes, as a function of the minimum underlying *P*-value threshold used to define the IGC callset (see Methods for IGC *P*-value calculation). Dashed lines represent individual haplotypes and the solid line represents the average. **d**, Correlation between IGC length and the number of supporting SNVs. **e**, Distribution of the distance between predicted IGC acceptor and donor sites for intrachromosomal events by chromosome.

Across all 102 haplotypes, we observe 121,631 putative IGC events for an average of 1,193 events per human haplotype (Fig. 2b,c and Supplementary Table 4). Of these events, 17,949 are rare and restricted to a single haplotype (singletons) whereas the remaining events are observed in several human haplotypes grouping into 14,663 distinct events (50% reciprocal overlap at both the donor and acceptor site). In total, we estimate that there is evidence for 32,612 different putative IGC events (Supplementary Table 5) among the SD regions that are assessed at present. Considering the redundant IGC callset ($n = 121{,}631$), the average IGC length observed in our data is 6.26 kbp with the largest event observed being 504 kbp (Extended Data Fig. 5). On average, each IGC event has 13.3 SNVs that support the conversion event and 2.03 supporting SNVs per kilobase pair, and as expected, there is strong correlation (Pearson's $R = 0.63$, $P < 2.2 \times 10^{-16}$; Fig. 2d) between the length of the events and supporting SNVs. Furthermore, we validated these supporting SNVs against Illumina sequencing data and find that on average

only 1% (12/1,192) of IGC events contain even one erroneous SNV (Supplementary Fig. 4). The putative IGC events detected with our method are largely restricted to higher identity duplications with only 325 events detected in 66.1 Mbp of SDs with >10% sequence divergence (Supplementary Figs. 14 and 15). We further stratify these results by callset, minimum number of supporting SNVs and haplotype (Supplementary Table 6). Finally, we use the number of supporting informative SNVs to estimate the statistical confidence of every putative IGC call (Fig. 2c, Supplementary Table 7 and Methods). Using these $P$ values, we identify a subset of the high-confidence ($P$ value < 0.05) IGC calls with 31,910 IGC events and 10,102 nonredundant events.

On average, we identify 7.5 Mbp of sequence per haplotype affected by putative IGC and 4.3 Mbp in our high-confidence callset (Fig. 2b). Overall, 33.8% (60.77/180.0 Mbp) of the analysed SD sequence is affected by putative IGC in at least one human haplotype. Furthermore, among all SDs covered by at least 20 assembled haplotypes, we identify 498 acceptor and 454 donor IGC hotspots with at least 20 distinct IGC events (Fig. 3 and Supplementary Table 8). IGC hotspots are more likely to associate with higher copy number SDs compared to a random sample of SD windows of equal size (median of 9 overlaps compared to 3, one-sided Wilcoxon rank sum test $P < 2.2 \times 10^{-16}$) and regions with more IGC events are moderately correlated with the copy number of the SD (Pearson's $R = 0.23$, $P < 2.2 \times 10^{-16}$; Supplementary Fig. 16). IGC hotspots also preferentially overlap higher identity duplications (median 99.4%) compared to randomly sampled windows (median 98.0%, one-sided Wilcoxon rank sum test $P < 2.2 \times 10^{-16}$).

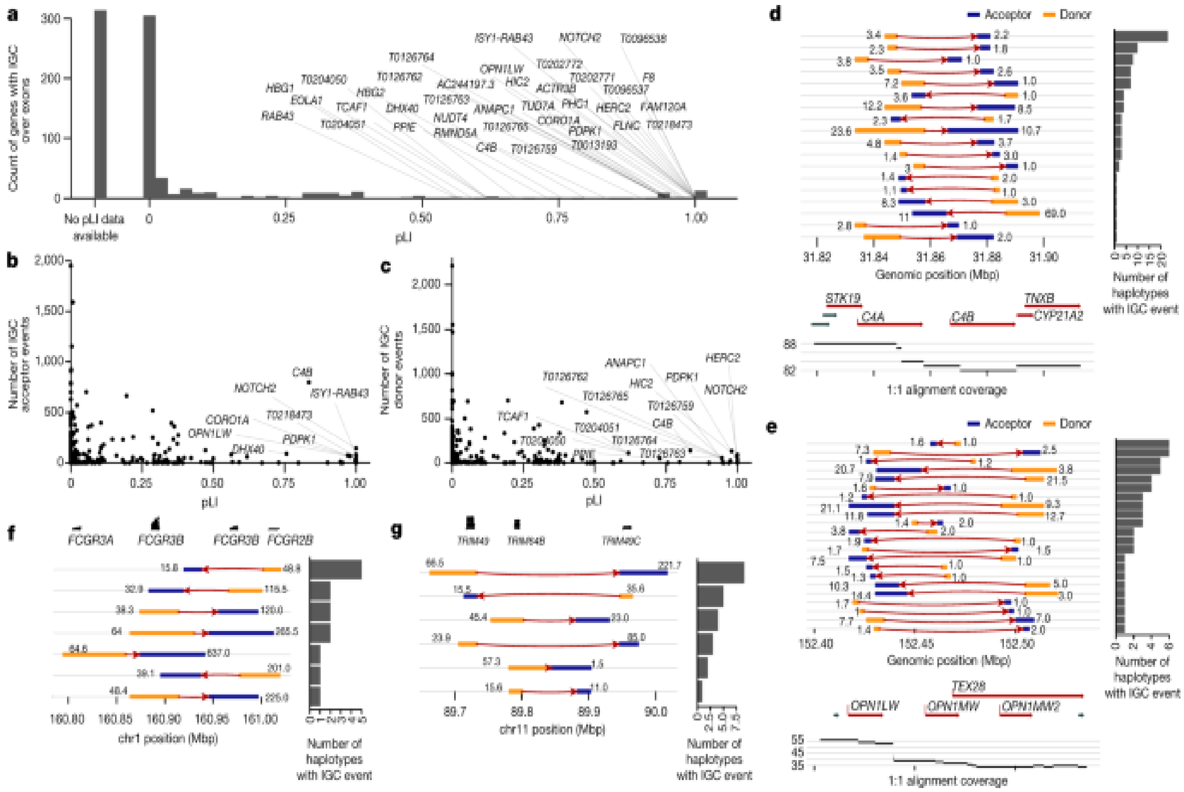
**Fig. 3: IGC hotspots.**



**a**, Density of IGC acceptor (top, blue) and donor (bottom, orange) sites across the ‘SD genome’. The SD genome consists of all main SD regions (>50 kbp) minus the intervening unique sequences. **b**, All intrachromosomal IGC events on 24 human haplotypes analysed for chromosome 15. Arcs drawn in blue (top) have the acceptor site on the left-hand side and the donor site on the right. Arcs drawn in orange (bottom) are arranged oppositely. Protein-coding genes are drawn as vertical black lines above the ideogram, and large duplication (blue) and deletion (red) events associated with human diseases are drawn as horizontal lines just above the ideogram. **c**, Zoom of the 30 highest confidence (lowest *P* value) IGC events on chromosome 15 between 17 and 31 Mbp. The number to the left of each event shows its length (kbp) and that to the right shows its number of SNVs. Genes with IGC events are highlighted in red and associate with the breakpoint regions of Prader–Willi syndrome. An expanded graphic with all haplotypes is included in Extended Data Fig. 7.

These events intersect 1,179 protein-coding genes, and of these genes, 799 have at least one coding exon affected by IGC (Supplementary Tables 9 and 10). As a measure of functional constraint, we used the probability of being loss-of-function intolerant (pLI) for each of the 799 genes[34] (Fig. 4a). Among these, 314 (39.3%) have never been assessed for mutation intolerance (that is, no pLI) owing to the limitations of mapping short-read data from population samples[34]. Of the remaining genes, we identify 38 with a pLI greater than 0.5, including genes associated with disease (*F8*, *HBG1* and *C4B*) and human evolution (*NOTCH2* and *TCAF*). Of the genes with high pLI scores, 12 are the acceptor site for at least 50 IGC events, including *CB4*, *NOTCH2* and *OPNL1W*—a locus for red–green colour blindness (Fig. 4b–e). We identify a subset of 418 nonredundant IGC events that are predicted to copy the entirety of a gene body to a 'new location' in the genome (Fig. 4f,g). As a result, 171 different protein-coding genes with at least 2 exons and 200 coding base pairs are converted in their entirety by putative IGC events in a subset of human haplotypes (Supplementary Table 11), and we refer to this phenomenon as gene repositioning. These gene-repositioning events are large (average 26 kbp; median 16.7 kbp) and supported by a high number of SNVs (average 64.7; median 15.3 SNVs), suggesting that they are unlikely to be mapping artefacts. Markedly, these putative IGC events copy the reference gene model on average a distance of 1.66 Mbp (median 216 kbp) from its original location. These include several disease-associated genes (for example, *TAOK2*, *C4A*, *C4B*, *PDPK1* and *IL27*) as well as genes that have eluded complete characterization owing to their duplicative nature[35,36,37].

**Fig. 4: Protein-coding genes affected by IGC.**




**a**, Number of punitive IGC events intersecting exons of protein-coding genes as a function of a gene's pLI. Of the 799 genes, 314 (39.3%) did not have a pLI score and are shown in the column labelled No pLI data available. **b**,**c**, Number of times a gene exon acts as an acceptor (**b**) or a donor (**c**) of an IGC event. **d**,**e**, IGC events at the complement factor locus, *C4A* and *C4B* (**d**), and the opsin middle- and long-wavelength-sensitive genes associated with colour blindness (*OPN1MW* and *OPN1LW* locus; **e**). Predicted donor (orange) and acceptor (blue) segments by length (number to left of event) and average number of supporting SNVs (number to right of event) are shown. The number of human haplotypes supporting each configuration is depicted by the histograms to the right. **f**,**g**, IGC events that reposition entire gene models for the *FCGR* (**f**) and *TRIM* (**g**) loci.

# Evolutionary age of SDs

Our analysis suggests that putative IGC contributes modestly to the significant increase of human SNV diversity in SDs. For example, if we apply the least conservative definition of IGC (1 supporting SNV) and exclude all putative IGC events from the human haplotypes, we estimate that

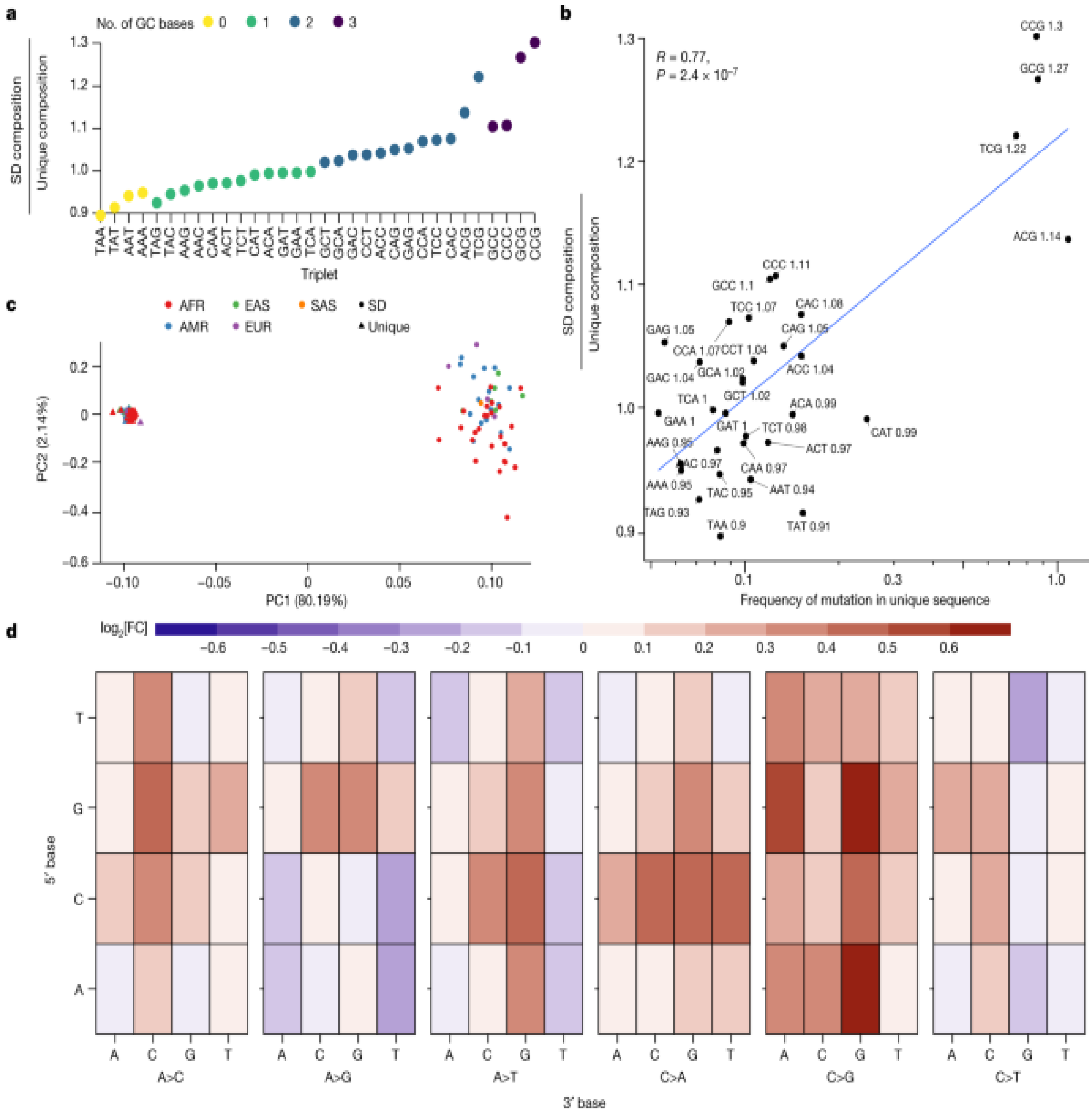
it accounts for only 23% of the increase (Extended Data Fig. 6). If we restrict to higher confidence IGC events ($P < 0.05$), only 19.6% of the increase could be accounted for. An alternative explanation may be that the SDs are evolutionarily older, perhaps owing to reduced selective constraint on duplicated copies[38,39]. To test whether SD sequences seem to have a deeper average coalescence than unique regions, we constructed a high-quality, locally phased assembly (hifiasm v0.15.2) of a chimpanzee (*Pan troglodytes*) genome to calibrate age since the time of divergence and to distinguish ancestral versus derived alleles in human SD regions (Methods). Constraining our analysis to syntenic regions between human and chimpanzee genomes (Methods), we characterized 4,316 SD regions (10 kbp in size) where we had variant calls from at least 50 human and one chimpanzee haplotype. We selected at random 9,247 analogous windows from unique regions for comparison. We constructed a multiple sequence alignment for each window and estimated the time to the most recent common ancestor (TMRCA) for each 10-kbp window independently. We infer that SDs are significantly older than the corresponding unique regions of similar size (Supplementary Figs. 17 and 18; one-sided Wilcoxon rank sum test $P$ value = $4.3 \times 10^{-14}$), assuming that mutation rates have remained constant over time within these regions since the human–chimpanzee divergence. The TMRCAs inferred from SD regions are, on average, 22% more ancient when compared to unique regions (650 versus 530 thousand years ago (ka)), but only a 5% difference is noted when comparing the median (520 versus 490 ka). However, this effect all but disappears (only a 0.2% increase) after excluding windows classified as IGC (Supplementary Fig. 19; one-sided Wilcoxon rank sum test $P = 0.05$; mean $TMRCA_{unique}$ = 528 ka, mean $TMRCA_{SD}$ = 581 ka, median $TMRCA_{unique}$ = 495 ka, median $TMRCA_{SD}$ = 496 ka).

## SNV mutational spectra in SDs

As a third possibility, we considered potential differences in the sequence context of unique and duplicated DNA. It has been recognized for almost two decades that human SDs are particularly biased towards Alu repeats and GC-rich DNA of the human genome[16,40]. Notably, among the SNVs in SDs, we observed a significant excess of transversions (transition/transversion

ratio (Ti/Tv) = 1.78) when compared to unique sequence (Ti/Tv = 2.06; $P < 2.2 \times 10^{-16}$, Pearson's chi-squared test with Yates's continuity correction). Increased mutability of GC-rich DNA is expected and may explain, in part, the increased variation in SDs and transversion bias[6,27,41]. Using a more complete genome, we compared the GC composition of unique and duplicated DNA specifically for the regions considered in this analysis. We find that, on average, 42.4% of the analysed SD regions are guanine or cytosine (43.0% across all SDs) when compared to 40.8% of the unique DNA ($P$ value $< 2.2 \times 10^{-16}$, one-sided $t$-test). Notably, this enrichment drops slightly (41.8%) if we exclude IGC regions. Consequently, we observe an increase of all GC-containing triplets in SD sequences compared to unique regions of the genome (Fig. 5a). Furthermore, the enrichment levels of particular triplet contexts in SD sequence correlate with the mutability of the same triplet sequence in unique regions of the genome (Pearson's $R = 0.77$, $P = 2.4 \times 10^{-7}$; Fig. 5b). This effect is primarily driven by CpG-containing triplets, which are enriched between 14 and 30% in SD sequences. Note, we observe a weaker and insignificant correlation for the non-CpG-containing triplets (Pearson's $R = 0.22$, $P = 0.27$). Extrapolating from the mutational frequencies seen in unique sequences, we estimate that there is 3.21% more variation with SDs due to their sequence composition alone.

**Fig. 5: Sequence composition and mutational spectra of SD SNVs.**




**a**, Compositional increase in GC-containing triplets in SD versus unique regions of the genome (coloured by GC content). **b**, Correlation between the enrichment of certain triplets in SDs compared to the mutability of that triplet in unique regions of the genome. Mutability is defined as the sum of all SNVs that change a triplet divided by the total count of that triplet in the genome. The enrichment ratio of SD over unique regions is indicated in text next to each triplet sequence. The text (upper left) indicates the value of the Pearson's correlation coefficient and the $P$ value from a two-sided $t$-test without adjustment for multiple comparisons. **c**, PCA of the mutational spectra of triplets in SD (circles) versus unique (triangles) regions polarized

against a chimpanzee genome assembly and coloured by the continental superpopulation of the sample. AFR, African; AMR, American; EAS, East Asian; EUR, European; SAS, South Asian. **d**, The log[fold change] in triplet mutation frequency between SD and unique sequences. The *y* axis represents the 5′ base of the triplet context; the first level of the *x* axis shows which central base has changed and the second level of the *x* axis shows the 3′ base: heatmap depicts the log[fold change]. As an example, the top left corner shows the log[fold change] in frequency of TAA>TCA mutations in SD versus unique sequences.

To further investigate the changes in GC content and their effect on variation in SDs, we compared the triplet mutational spectra of SNVs from unique and duplicated regions of the genome to determine whether the predominant modes of SNV mutation differed (Methods). We considered all possible triplet changes, first quantifying the number of ancestral GC bases and triplets in SDs (Fig. 5a). A principal component analysis (PCA) of these normalized mutational spectra shows clear discrimination (Fig. 5c) between unique and SD regions (PC1) beyond that of African and non-African diversity, with the first principal component capturing 80.2% of the variation separating the mutational spectrum of SDs and unique DNA. We observe several differences when comparing the triplet-normalized mutation frequency of particular mutational events in SD and unique sequences (Fig. 5d). Most notable is a 7.6% reduction in CpG transition mutations—the most predominant mode of mutation in unique regions of the genome due to spontaneous deamination of methylated CpGs[6] (Supplementary Tables 12 and 13).

The most notable changes in mutational spectra in SD sequences are a 27.1% increase in C>G mutations, a 15.3% increase in C>A mutations and a 10.5% increase in A>C mutations. C>G mutations are associated with double-strand breaks in humans and some other apes[42,43]. This effect becomes more pronounced (+40.4%) in our candidate IGC regions consistent with previous observations showing increases in C>G mutations in regions of non-crossover gene conversion and double-strand breaks[43,44,45]. However, the increase remains in SD regions without IGC (+20.0%) perhaps owing to extensive nonallelic homologous recombination associated with SDs or undetected IGC events[4,9].

To further investigate the potential effect of GC-biased gene conversion (gBGC) on the mutational spectra in SDs, we measured the frequency of (A,T)>(G,C) mutations in SD regions with evidence of IGC to determine whether cytosine and guanine bases are being preferentially maintained as might be expected in regions undergoing gBGC. If we measure the frequency of (A,T)>(C,G) in windows with at least one haplotype showing evidence of IGC, then we observe that the frequency is 4.7% higher than in unique regions of the genome; notably, in SDs without IGC, this rate is reduced compared to that of unique sequence (−3.5%). Additionally, there is a 5.8% reduction in (G,C)>(A,T) bases consistent with IGC preferentially restoring CG bases that have mutated to AT bases through gBGC. These results indicate that gBGC between paralogous sequences may be a strong factor in shaping the mutational landscape of SDs. Although, the (A,T)>(C,G) frequency is comparable in SD regions not affected by IGC, the mutational landscape at large is still very distinct between SDs and unique parts of the genome. In PCA of the mutational spectra in SDs without IGC, the first principal component distinguishing the mutational spectrum of SDs and unique DNA captures a larger fraction of the variation (94.6%) than in the PCA including IGC sites (80.2%; Supplementary Fig. 20).

## Modelling of elevated SNV frequency

To model the combined effect of unique mutational properties, evolutionary age and sequence content on the frequency of SNVs, we developed a multivariable linear regression using copy number, SD identity, number of unique IGC events, GC content and TMRCA to predict the number of SNVs seen in a 10-kbp window. A linear model containing all pairwise interactions of these predictors was able to explain 10.5% of the variation in SNVs per 10 kbp (adjusted $R^2$), whereas a model containing only the number of IGC events explained only 1.8% of the variation. We note that this measure of variance is related but not directly comparable to the finding that the elevation in the number of SNVs is reduced by 23% when excluding IGC regions. All of the random variables, including their pairwise interactions, were significant ($P$ value < 0.05) predictors of SNVs per 10 kbp except the interaction of number of IGC events with GC content, copy number and TMRCA. The strongest single predictors were the number of unique IGC events and the divergence of the overlapping SD (Supplementary Table 14).

# Discussion

Since the first publications of the human genome[12,13], the pattern of single-nucleotide variation in recently duplicated sequence has been difficult to ascertain, leading to errors[2,11]. Later, indirect approaches were used to infer true SNVs in SDs, but these were far from complete[40]. More often than not, large-scale sequencing efforts simply excluded such regions in an effort to prevent paralogous sequence variants from contaminating single-nucleotide polymorphism databases and leading to false genetic associations[8,23]. The use of phased genome assemblies as opposed to aligned sequence reads had the advantage of allowing us to establish 1:1 orthologous relationships as well as the ability to discern the effect of IGC while comparing the pattern of single-nucleotide variation for both duplicated and unique DNA within the same haplotypes. As a result, we identify over 1.99 million nonredundant SNVs in a gene-rich portion of the genome previously considered largely inaccessible.

SNV density is significantly elevated (60%) in duplicated DNA when compared to unique DNA consistent with suggestions from primate genome comparisons and more recent de novo mutation studies from long-read sequencing data[46,47,48]. Furthermore, an increased de novo mutation rate in SDs could support our observation of an elevated SNV density without the need for an increase in TMRCA. We estimate that at least 23% of this increase is due to the action of IGC between paralogous sequences that essentially diversify allelic copies through concerted evolution. IGC in SDs seems to be more pervasive in the human genome compared to earlier estimates[15,27], which owing to mapping uncertainties or gaps could assay only a smaller subset of regions[15,27]. We estimate more than 32,000 candidate regions (including 799 protein-coding genes) with the average human haplotype showing 1,192 events when compared to the reference. The putative IGC events are also much larger (mean 6.26 kbp) than those of most previous reports[28,49], with the top 10% of the size distribution >14.4 kbp in length. This has the net effect that entire genes are copied hundreds of kilobase pairs into a new genomic context when compared to the reference. The effect of such ‘repositioning events’ on gene regulation will be an interesting avenue of future research.

As for allelic gene conversion, our predicted nonallelic gene conversion events are abundant, cluster into larger regional hotspots and favour G and C mutations, although this last property is not restricted to IGC regions[45,50]. Although we classify these regions as putative IGC events, other mutational processes such as deletion followed by duplicative transposition could, in principle, generate the same signal creating large tracts of ‘repositioned’ DNA. It should also be stressed that our method simply relies on the discovery of a closer match within the reference; by definition, this limits the detection of IGC events to regions where the donor sequence is already present in the reference as opposed to an alternative. Moreover, we interrogated only regions where 1:1 synteny could be unambiguously established. As more of the genome is assessed in the context of a pangenome reference framework, we anticipate that the proportion of IGC will increase, especially as large-copy-number polymorphic SDs, centromeres and acrocentric DNA become fully sequence resolved[3]. Although we estimate 4.3 Mbp of IGC in SDs on average per human haplotype, we caution that this almost certainly represents a lower bound and should not yet be regarded as a rate until more of the genome is surveyed and studies are carried out in the context of parent–child trios to observe germline events.

One of the most notable features of duplicated DNA is its higher GC content. In this study, we show that there is a clear skew in the mutational spectrum of SNVs to maintain this property of SDs beyond expectations from unique DNA. This property and the unexpected Ti/Tv ratio cannot be explained by lower accuracy of the assembly of SD regions. We find a 27.1% increase in transversions that convert cytosine to guanine or the reverse across all triplet contexts. GC-rich DNA has long been regarded as hypermutable. For example, C>G mutations preferentially associate with double-strand breaks in humans and apes[42,43] and GC-rich regions in yeast show about 2–5 times more mutations depending on sequence context compared to AT-rich DNA[41]. Notably, in human SD regions, we observe a paucity of CpG transition mutations, characteristically associated with spontaneous deamination of CpG dinucleotides and concomitant transitions[6]. The basis for this is unclear, but it may be partially explained by the recent observation that duplicated genes show a greater degree of hypomethylation when compared to their unique counterparts[10]. We propose

that excess of guanosine and cytosine transversions is a direct consequence of GC-biased gene conversion[5] driven by an excess of double-strand breaks that result from a high rate of nonallelic homologous recombination events and other break-induced replication mechanisms among paralogous sequences.

# Methods

## Defining unique and SD regions

To define regions of SD, we used the annotations available for T2T-CHM13 v1.1 (ref. [10]), which include all nonallelic intrachromosomal and interchromosomal pairwise alignments >1 kbp and with >90% sequence identity that do not consist entirely of common repeats or satellite sequences[11]. To define unique regions, we found the coordinates in T2T-CHM13 that were not SDs, ancient SDs (<90% sequence identity), centromeres or satellite arrays[51] and defined these areas to be the non-duplicated (unique) parts of the genome. For both SDs and unique regions, variants in tandem repeat elements as identified by Tandem Repeats Finder[52] were excluded because many SNVs called in these regions are ultimately alignment artefacts. RepeatMasker v4.1.2 was used to annotate SNVs with additional repeat classes beyond SDs[53].

## Copy number estimate validation

The goal of this analysis was to validate copy number from the assembled HPRC haplotypes compared to estimates from read-depth analysis of the same samples sequenced using Illumina whole-genome sequencing (WGS). Large, recently duplicated segments are prone to copy number variation and are also susceptible to collapse and misassembly owing to their repetitive nature. HPRC haplotypes were assembled using PacBio HiFi with hifiasm[3,54] creating contiguous long-read assemblies. We selected 19 SD loci corresponding to genes that were known to be duplicated and copy number variable in the human species. We *k*-merized the 2 haplotype assemblies corresponding to each locus for each individual into *k*-mers of 31 base pairs in length. We then computed copy number estimates over each

locus for the sum haplotype assemblies and calculated the difference based on Illumina WGS from the same sample. For both datasets, we derived these estimates using FastCN, an algorithm implementing whole-genome shotgun sequence detection[55]. When averaging across each region and comparing differences in assembly copy versus Illumina WGS copy estimate, we observe that 756 out of 893 tests were perfectly matched ($\delta = 0$), suggesting that most of these assemblies correctly represent the underlying genomic sequence of the samples.

## Quality value estimations with Merqury

Estimates of the quality value of SD and unique regions were made using Merqury v1.1 and parental Illumina sequencing data[56]. We first used Meryl to create *k*-mer databases (with a *k*-mer length of 21) using the parental sequencing data following the instructions in the Merqury documentation. Then Merqury was run with default parameters (merqury.sh {k-mer meryl database} {paternal sequence} {maternal sequence}) to generate quality value estimates for the hifiasm assemblies.

## Haplotype integrity analysis using inter-SUNK approach

For the 35 HPRC assemblies with matched ultralong Oxford Nanopore Technologies (ONT) data, we applied GAVISUNK v1.0.0 as an orthogonal validation of HiFi assembly integrity[57]. In brief, candidate haplotype-specific singly unique nucleotide *k*-mers (SUNKs) of length 20 are determined from the HiFi assembly and compared to ONT reads phased with parental Illumina data. Inter-SUNK distances are required to be consistent between the assembly and ONT reads, and regions that can be spanned and tiled with consistent ONT reads are considered validated. ONT read dropouts do not necessarily correspond to misassembly—they are also caused by large regions devoid of haplotype-specific SUNKs from recent duplications, homozygosity or over-assembly of the region, as well as Poisson dropout of read coverage.

## Read-depth analysis using the HPRC unreliable callset

For the 94 assembled HPRC haplotypes, we downloaded the regions identified to have abnormal coverage form S3 (s3://human-pangenomics/submissions/e9ad8022-1b30-11ec-ab04-0a13c5208311–COVERAGE_ANALYSIS_Y1_GENBANK/FLAGGER/JAN_09_2022/FINAL_HIFI_BASED/FLAGGER_HIFI_ASM_SIMPLIFIED_BEDS/ALL/). We then intersected these regions with the callable SD regions in each assembly to determine the number of collapsed, falsely duplicated and low-coverage base pairs in each assembly. The unreliable regions were determined by the HPRC using Flagger v0.1 (https://github.com/mobinasri/flagger/)[3].

## Whole-genome alignments and synteny definition

Whole-genome alignments were calculated against T2T-CHM13 v1.1 with a copy of GRCh38 chrY using minimap2 v2.24 (ref. [58]) with the parameters -a -x asm20–secondary=no -s 25000 -K 8G. The alignments were further processed with rustybam v0.1.29 (ref. [59]) using the subcommands trim-paf to remove redundant alignments in the query sequence and break-paf to split alignments on structural variants over 10 kbp. After these steps, the remaining alignments over 1 Mbp of continuously aligned sequence were defined to be syntenic. The software pipeline is available on GitHub at https://github.com/mrvollger/asm-to-reference-alignment/ (refs. [58,59,60,61,62,63,64,65,66,67]).

## Estimating the diversity of SNVs in SDs and unique sequences

When enumerating the number of SNVs, we count all pairwise differences between the haplotypes and the reference, counting events observed in multiple haplotypes multiple times. Therefore, except when otherwise indicated, we are referring to the total number of pairwise differences rather than the total number of nonredundant SNVs (number of segregation sites). The software pipeline is available on GitHub at https://github.com/mrvollger/sd-divergence (refs. [60,61,62,63,65,66,68]).

## Defining IGC events

Each query haplotype genome sequence was aligned to the reference genome (T2T-CHM13 v1.1) using minimap2 v2.24 (ref. [58]) considering only those regions that align in a 1:1 fashion for >1 Mbp without any evidence of gaps or discontinuities greater than 10 kbp in size. This eliminates large forms of structural variation, including copy number variants or regions of large-scale inversion restricting the analysis to largely copy number invariant SD regions (about 120 Mbp) and flanking unique sequence. Once these syntenic alignments were defined, we carried out a second alignment fragmenting the 1:1 synteny blocks into 1-kbp windows (100-bp increments) and remapped back to T2T-CHM13 to identify each window's single best alignment position. These second alignments were then compared to original syntenic ones and if they no longer overlapped, we considered them to be candidate IGC regions. Adjacent IGC windows were subsequently merged into larger intervals when windows continued to be mapped non-syntenically with respect to the original alignment. We then used the CIGAR string to identify the number of matching and mismatching bases at the 'donor' site and compared that to the number of matching and mismatching bases at the acceptor site determined by the syntenic alignment. A donor sequence is, thus, defined as a segment in T2T-CHM13 that now maps with higher sequence identity to a new location in the human haplotype (alignment method 2) and the acceptor sequence is the segment in T2T-CHM13 that has an orthologous mapping to the same region in the human haplotype (alignment method 1). As such, there is dependence on both the reference genome and the haplotype being compared. The software pipeline is available on GitHub at https://github.com/mrvollger/asm-to-reference-alignment/ (refs. [58,59,60,61,62,63,64,65,66,67]).

## Assigning confidence to IGC events

To assign confidence measures to our IGC events, we adapted a previously described method[69] to calculate a *P* value for every one of our candidate IGC calls. Our method uses a cumulative binomial distribution constructed from the number of SNVs supporting the IGC event and the total number of informative sites between two paralogues to assign a one-sided *P* value to each event. Specifically:

$$P(X\le k)=B(k,n,p)$$

in which $B$ is the binomial cumulative distribution, $n$ is the number of informative sites between paralogues, $k$ is the number of informative sites that agree with the non-converted sequence (acceptor site), and $p$ is the probability that at an informative site the base matches the acceptor sequence. We assume $p$ to be 0.5 reflecting that a supporting base change can come from one of two sources: the donor or acceptor paralogue. With these assumptions, our binomial model reports the probability that we observe $k$ or fewer sites that support the acceptor site (that is, no IGC) at random given the data, giving us a one-sided $P$ value for each IGC event. No adjustments were made for multiple comparisons.

## Testing for IGC in unique regions

To test the specificity of our method, we applied it to an equivalent total of unique sequence (125 Mbp) on each haplotype, which we expected to show no or low levels of IGC. On average, we identify only 33.5 IGC events affecting 38.2 kbp of sequence per haplotype. If we restrict this to high-confidence IGC events, we see only 5.93 events on average affecting 7.29 kbp. This implies that our method is detecting IGC above background in SDs and that the frequency of IGC in SDs is more than 50 times higher in the high-confidence callsets (31,910 versus 605).

## Additional genome assemblies

We assembled HG00514, NA12878 and HG03125 using HiFi long-read data and hifiasm v0.15.2 with parental Illumina data[54]. Using HiFi long-read data and hifiasm v0.15.2 we also assembled the genome of the now-deceased chimpanzee Clint (sample S006007). The assembly is locally phased as trio-binning and HiC data were unavailable. Data are available on the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under the BioProjects PRJNA551670 (ref. [4]), PRJNA540705 (ref. [70]), PRJEB36100 (ref. [4]) and PRJNA659034 (ref. [47]). These assemblies are made available on Zenodo (https://doi.org/10.5281/zenodo.6792653)[71].

## Determining the composition of triplet mutations in SD and unique sequences

The mutational spectra for unique and SD regions from each individual were computed using mutyper on the basis of derived SNVs polarized against the chimpanzee genome assembly described above[72,73,74]. These spectra were normalized to the triplet content of the respective unique or SD regions by dividing the count of each triplet mutation type by the total count of each triplet context in the ancestral region and normalizing the number of counts in SD and unique sequences to be the same. For PCA, the data were further normalized using the centred log-ratio transformation, which is commonly used for compositional measurements[75]. The code is available on GitHub at https://github.com/mrvollger/mutyper_workflow/ (refs. [61,62,63,65,72,76]).

### Estimation of TMRCA

To estimate TMRCA for a locus of interest, we focus on orthologous sequences (10-kbp windows) identified in synteny among human and chimpanzee haplotypes. Under an assumption of infinite sites, the number of mutations $x_i$ between a human sequence and its most recent common ancestor is Poisson distributed with a mean of $\mu \times T$, in which $\mu$ is the mutation rate scaled with respect to the substitutions between human and chimpanzee lineages, and *T* is the TMRCA. That is, $T=\sum_{i=1}^{n} x_i/n\mu$, in which *n* is the number of human haplotypes. To convert TMRCA to time in years, we assume six million years of divergence between human and chimpanzee lineages. We note that the TMRCA estimates reported in the present study account for mutation variation across loci (that is, if the mutation rate is elevated for a locus, the effect would be accounted for). Thus, for each individual locus, an independent mutation (not uniform) rate is applied depending on the observed pattern of mutations compared to the chimpanzee outgroup.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

PacBio HiFi and ONT data have been deposited into NCBI SRA under the following BioProject IDs: PRJNA850430, PRJNA731524, PRJNA551670, PRJNA540705 and PRJEB36100. PacBio HiFi data for CHM1 are available under the following SRA accessions: SRX10759865 and SRX10759866. Sequencing data for Clint PTR are available on NCBI SRA under the BioProject PRJNA659034. The T2T-CHM13 v1.1 assembly can be found on NCBI (GCA_009914755.3). Cell lines obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research are listed in Supplementary Table 1. Assemblies of HPRC samples are available on NCBI under the BioProject PRJNA730822. All additional assemblies used in this work (Clint PTR, CHM1, HG00514, NA12878 and HG03125), variant calls, assembly alignments, and other annotation data used in analysis are available on Zenodo (https://doi.org/10.5281/zenodo.6792653)[71].

# Code availability

The software pipeline for aligning assemblies and calling IGC is available on GitHub (https://github.com/mrvollger/asm-to-reference-alignmentv0.1) and Zenodo (https://zenodo.org/record/7653446)[67]. Code for analysing variants called against T2T-CHM13 v1.1 is available on GitHub (https://github.com/mrvollger/sd-divergencev0.1 and Zenodo (https://zenodo.org/record/7653464)[68]. The software pipeline for analysing the triple context of SNVs is available on GitHub (https://github.com/mrvollger/mutyper_workflowv0.1) and Zenodo (https://zenodo.org/record/7653472)[76]. Scripts for figure and table generation are available on GitHub (https://github.com/mrvollger/sd-divergence-and-igc-figuresv0.1) and Zenodo (https://zenodo.org/record/7653486)[77]. GAVISUNK is available on GitHub (https://github.com/pdishuck/GAVISUNK) and Zenodo (https://zenodo.org/record/7655335)[57].

# References


1. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current

human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).

2. Fredman, D. et al. Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet.* **36**, 861–866 (2004).

3. Liao, W.-W. et al. A draft human pangenome reference. *Nature*, https://doi.org/10.1038/s41586-023-05896-x (2023).

4. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

5. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).

6. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).

7. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).

8. 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

9. Sudmant, P. H. et al. Diversity of human copy number. *Science* **11184**, 2–7 (2010).

10. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).

11. Bailey, J. A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).

12. IHGSC. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

13. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

14. Sharp, A. J. et al. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).

15. Dumont, B. L. Interlocus gene conversion explains at least 2.7% of single nucleotide variants in human segmental duplications. *BMC Genomics* **16**, 456 (2015).

16. Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834 (2003).

17. Jiang, Z. et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat. Genet.* **39**, 1361–1368 (2007).

18. Nuttle, X. Emergence of a *Homo sapiens*-specific gene family and chromosome 16p11. 2 CNV susceptibility. *Nature* **536**, 205–209 (2016).

19. Dougherty, M. L. et al. Transcriptional fates of human-specific segmental duplications in brain. *Genome Res.* **28**, 1566–1576 (2018).

20. Fiddes, I. T. et al. Human-specific *NOTCH2NL* genes affect notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369 (2018).

21. Ju, X.-C. et al. The hominoid-specific gene *TBC1D3* promotes generation of basal neural progenitors and induces cortical folding in mice. *eLife* **5**, e18197 (2016).

22. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).

23. Zook, J. M. et al. An open resource for accurately benchmarking small variant and reference calls. *Nat. Biotechnol.* **37**, 561–566 (2019).

24. Teshima, K. M. & Innan, H. The coalescent with selection on copy number variants. *Genetics* **190**, 1077–1086 (2012).

25. Innan, H. The coalescent and infinite-site model of a small multigene family. *Genetics* **163**, 803–810 (2003).

26. Hartasánchez, D. A., Vallès-Codina, O., Brasó-Vives, M. & Navarro, A. Interplay of interlocus gene conversion and crossover in segmental duplications under a neutral scenario. *G3 Genes Genomes Genet.* **4**, 1479–1489 (2014).

27. Harpak, A., Lan, X., Gao, Z. & Pritchard, J. K. Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proc. Natl Acad. Sci. USA* **114**, 201708151 (2017).

28. Mansai, S. P., Kado, T. & Innan, H. The rate and tract length of gene conversion between duplicated genes. *Genes* **2**, 313–331 (2011).

29. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

30. Jarvis, E. D. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).

31. Porubsky, D. et al. Gaps and complex structurally variant loci in phased genome assemblies. *Genom. Res.* https://doi.org/10.1101/gr.277334.122 (2023).

32. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01662-6 (2023).

33. Bosch, E., Hurles, M. E., Navarro, A. & Jobling, M. A. Dynamics of a human interparalog gene conversion hotspot. *Genome Res.* **14**, 835–844 (2004).

34. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

35. Richter, M. et al. Altered TAOK2 activity causes autism-related neurodevelopmental and cognitive abnormalities through RhoA signaling. *Mol. Psychiatry* **24**, 1329–1350 (2019).

36. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).

37. Pietri, M. et al. PDK1 decreases TACE-mediated α-secretase activity and promotes disease progression in prion and Alzheimer's diseases. *Nat. Med.* **19**, 1124–1131 (2013).

38. Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).

39. Conant, G. C. & Wagner, A. Asymmetric sequence divergence of duplicate genes. *Genome Res.* **13**, 2052–2058 (2003).

40. Nakken, S., Rødland, E. A., Rognes, T. & Hovig, E. Large-scale inference of the point mutational spectrum in human segmental duplications. *BMC Genomics* **10**, 43 (2009).

41. Kiktev, D. A., Sheng, Z., Lobachev, K. S. & Petes, T. D. GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **115**, E7109–E7118 (2018).

42. Goldmann, J. M. et al. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).

43. Gao, Z. et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl Acad. Sci. USA* **116**, 9491–9500 (2019).

44. Elliott, B., Richardson, C., Winderbaum, J., Nickoloff, J. A. & Jasin, M. Gene conversion tracts from double-strand break repair in mammalian cells. *Mol. Cell. Biol.* **18**, 93–101 (1998).

45. Williams, A. L. et al. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* **4**, e04637 (2015).

46. Liu, G. et al. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13**, 358–368 (2003).

47. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).

48. Noyes, M. D. et al. Familial long-read sequencing increases yield of de novo mutations. *Am. J. Hum. Genet.* **109**, 631–646 (2022).

49. Ji, X. & Thorne, J. L. A phylogenetic approach disentangles interlocus gene conversion tract length and initiation rate. Preprint at https://arxiv.org/abs/1908.08608 (2019).

50. Narasimhan, V. M. et al. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat. Commun.* **8**, 303 (2017).

51. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).

52. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

53. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0, http://www.repeatmasker.org (2013–2015).

54. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

55. Pendleton, A. L. et al. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.* **16**, 64 (2018).

56. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

57. Dishuck, P. C., Rozanski, A. N., Logsdon, G. A., Porubsky, D. & Eichler, E. E. GAVISUNK: genome assembly validation via inter-SUNK distances in Oxford Nanopore reads. *Bioinformatics* **39**, btac714 (2022).

58. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

59. Vollger, M. R. mrvollger/rustybam: v0.1.29. *Zenodo*, https://doi.org/10.5281/ZENODO.6342176. (2022)

60. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

61. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

62. Bonfield, J. K. et al. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**, giab007 (2021).

63. Mölder, F. et al. Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).

64. pysam: a Python module for reading and manipulating SAM/BAM/VCF/BCF files. *GitHub*, https://github.com/pysam-developers/pysam (2021).

65. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1-34 (2014).

66. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).

67. Vollger, M. R. mrvollger/asm-to-reference-alignment: v0.1. *Zenodo*, https://doi.org/10.5281/ZENODO.7653446 (2023).

68. Vollger, M. R. mrvollger/sd-divergence: v0.1. *Zenodo*, https://doi.org/10.5281/ZENODO.7653464 (2023).

69. Carey, K. M., Patterson, G. & Wheeler, T. J. Transposable element subfamily annotation has a reproducibility problem. *Mob. DNA* **12**, 4 (2021).

70. Porubsky, D. et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.* **39**, 302–308 (2021).

71. Vollger, M. Supplementary data for: Increased mutation and gene conversion within human segmental duplications. *Zenodo*, https://doi.org/10.5281/zenodo.7651064 (2023).

72. DeWitt, W. S. mutyper: assigning and summarizing mutation types for analyzing germline mutation spectra. Preprint at https://doi.org/10.1101/2020.07.01.183392 (2020).

73. Carlson, J., DeWitt, W. S. & Harris, K. Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation. *Curr. Opin. Genet. Dev.* **62**, 50–57 (2020).

74. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl Acad. Sci. USA* **112**, 3439–3444 (2015).

75. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc.* **44**, 139–160 (1982).

76. Vollger, M. R. mrvollger/mutyper_workflow: v0.1. *Zenodo*, https://doi.org/10.5281/ZENODO.7653472 (2023).

77. Vollger, M. R. mrvollger/sd-divergence-and-igc-figures: v0.1. *Zenodo*, https://doi.org/10.5281/ZENODO.7653486 (2023).

# Acknowledgements

We thank T. Brown for help in editing this manuscript, P. Green for valuable suggestions, and R. Seroussi and his staff for their generous donation of time and resources. This work was supported in part by grants from the US National Institutes of Health (NIH 5R01HG002385, 5U01HG010971 and 1U01HG010973 to E.E.E.; K99HG011041 to P.H.; and F31AI150163 to W.S.D.). W.S.D. was supported in part by a Fellowship in Understanding Dynamic and Multi-scale Systems from the James S. McDonnell

Foundation. E.E.E. is an investigator of the Howard Hughes Medical Institute (HHMI). This article is subject to HHMI’s Open Access to Publications policy. HHMI laboratory heads have previously granted a nonexclusive CC BY 4.0 licence to the public and a sublicensable licence to HHMI in their research articles. Pursuant to those licences, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 licence immediately on publication.

# Author information

## Authors and Affiliations

1. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

   Mitchell R. Vollger, Philip C. Dishuck, William T. Harvey, William S. DeWitt, Xavi Guitart, Michael E. Goldberg, Allison N. Rozanski, Carl A. Baker, Jennifer Kordosky, Mitchell R. Vollger, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, Glennis A. Logsdon, David Porubsky, Kelley Harris, PingHsun Hsieh & Evan E. Eichler

2. Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA, USA

   Mitchell R. Vollger & Mitchell R. Vollger

3. Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

   William S. DeWitt

4. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA

   William S. DeWitt

5. UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA

Julian Lucas, Mobin Asri, Xian H. Chang, Mark Diekhans, Jordan M. Eizenga, Marina Haukness, David Haussler, Glenn Hickey, Julian K. Lucas, Charles Markello, Karen H. Miga, Jean Monlong, Adam M. Novak, Hugh E. Olsen, Benedict Paten, Trevor Pesout, Jouni Sirén & Benedict Paten

6. Howard Hughes Medical Institute, Chevy Chase, MD, USA

David Haussler, Erich D. Jarvis & Evan E. Eichler

7. Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St Louis, MO, USA

Haley J. Abel

8. McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO, USA

Lucinda L. Antonacci-Fulton, Sarah Cody, Robert S. Fulton, Allison A. Regier, Chad Tomlinson & Ting Wang

9. Google LLC, Mountain View, CA, USA

Gunjan Baid, Anastasiya Belyaeva, Andrew Carroll, Pi-Chuan Chang, Daniel E. Cook, Alexey Kolesnikov, Maria Nattestad & Kishwar Shafin

10. European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

Konstantinos Billis, Susan Fairley, Paul Flicek, Adam Frankish, Carlos Garcia Giron, Leanne Haggerty, Thibaut Hourlier, Jan O. Korbel, Fergal J. Martin & Francesca Floriana Tricomi

11. Department of Human Genetics, McGill University, Montreal, Quebec, Canada

Guillaume Bourque

12. Canadian Center for Computational Genomics, McGill University, Montreal, Quebec, Canada

    Guillaume Bourque

13. Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan

    Guillaume Bourque

14. Institute of Genetics and Biophysics, National Research Council, Naples, Italy

    Silvia Buonaiuto & Vincenza Colonna

15. Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA

    Mark J. P. Chaisson & Tsung-Yu Lu

16. Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

    Haoyu Cheng, Justin Chu, Xiaowen Feng & Heng Li

17. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

    Haoyu Cheng, Xiaowen Feng & Heng Li

18. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

    Vincenza Colonna, Christian Fischer, Erik Garrison, Andrea Guarracino, Pjotr Prins & Flavia Villani

19. Barrett and O’Connor Washington Center, Arizona State University, Washington DC, USA

Robert M. Cook-Deegan

20. Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA, USA

Omar E. Cornejo

21. Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Daniel Doerr, Peter Ebert, Jana Ebler, Hugo Magalhães, Pierre Marijon & Tobias Marschall

22. Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Daniel Doerr, Peter Ebert, Jana Ebler, Hugo Magalhães, Pierre Marijon & Tobias Marschall

23. Core Unit Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Peter Ebert

24. Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA

Olivier Fedrigo, Giulio Formenti, Erich D. Jarvis & Jacquelyn Mountcastle

25. National Institutes of Health (NIH)–National Human Genome Research Institute, Bethesda, MD, USA

Adam L. Felsenfeld, Baergen I. Schultz, Michael W. Smith & Heidi J. Sofia

26. Department of Genetics, Washington University School of Medicine, St Louis, MO, USA

Robert S. Fulton & Ting Wang

27. Center for Computational and Genomic Medicine, The Children’s Hospital of Philadelphia, Philadelphia, PA, USA

Yan Gao

28. Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Copenhagen, Denmark

Shilpa Garg

29. Institute for Society and Genetics, College of Letters and Science, University of California, Los Angeles, Los Angeles, CA, USA

Nanibaa’ A. Garrison

30. Institute for Precision Health, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

Nanibaa’ A. Garrison

31. Division of General Internal Medicine and Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

Nanibaa’ A. Garrison

32. Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA

Richard E. Green

33. Dovetail Genomics, Scotts Valley, CA, USA

Richard E. Green

34. Quantitative Life Sciences, McGill University, Montreal, Quebec, Canada

Cristian Groza

35. Genomics Research Centre, Human Technopole, Milan, Italy

Andrea Guarracino

36. Department of Genetics, Yale University School of Medicine, New Haven, CT, USA

Ira M. Hall, Wen-Wei Liao & Shuangjia Lu

37. Center for Genomic Health, Yale University School of Medicine, New Haven, CT, USA

Ira M. Hall & Wen-Wei Liao

38. Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany

Simon Heumos

39. Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen, Germany

Simon Heumos

40. Tree of Life, Wellcome Sanger Institute, Hinxton, UK

Kerstin Howe & Jonathan M. D. Wood

41. Northeastern University, Boston, MA, USA

Miten Jain

42. Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA

Erich D. Jarvis

43. Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

Hanlee P. Ji & HoJoon Lee

44. Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Eimear E. Kenny

45. Program in Bioethics and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA

Barbara A. Koenig

46. Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

Jan O. Korbel

47. Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Sergey Koren, Ann McCartney, Sergey Nurk, Adam M. Phillippy, Mikko Rautiainen, Arang Rhie & Brian Walenz

48. Division of Biology and Biomedical Sciences, Washington University School of Medicine, St Louis, MO, USA

Wen-Wei Liao

49. Computer Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

Santiago Marco-Sola

50. Departament d’Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Barcelona, Spain

Santiago Marco-Sola

51. Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

Jennifer McDaniel, Nathan D. Olson, Justin Wagner & Justin M. Zook

52. Coriell Institute for Medical Research, Camden, NJ, USA

Matthew W. Mitchell

53. Department of Computer Science, University of Pisa, Pisa, Italy

Moses Njagi Mwaniki

54. Department of Public Health Sciences, University of California, Davis, Davis, CA, USA

Alice B. Popejoy

55. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

Daniela Puiu & Aleksey V. Zimin

56. Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA, USA

Samuel Sacco

57. Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany

Ashley D. Sanders

58. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Valerie A. Schneider & Françoise Thibaud-Nissen

59. Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark

    Jonas A. Sibbesen

60. Al Jalila Genomics Center of Excellence, Al Jalila Children's Specialty Hospital, Dubai, United Arab Emirates

    Ahmad N. Abou Tayoun

61. Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates

    Ahmad N. Abou Tayoun

62. Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA

    Aleksey V. Zimin

## Consortia

## Human Pangenome Reference Consortium

- Haley J. Abel
- , Lucinda L. Antonacci-Fulton
- , Mobin Asri
- , Gunjan Baid
- , Carl A. Baker
- , Anastasiya Belyaeva
- , Konstantinos Billis
- , Guillaume Bourque
- , Silvia Buonaiuto
- , Andrew Carroll
- , Mark J. P. Chaisson
- , Pi-Chuan Chang

- , Xian H. Chang
- , Haoyu Cheng
- , Justin Chu
- , Sarah Cody
- , Vincenza Colonna
- , Daniel E. Cook
- , Robert M. Cook-Deegan
- , Omar E. Cornejo
- , Mark Diekhans
- , Daniel Doerr
- , Peter Ebert
- , Jana Ebler
- , Evan E. Eichler
- , Jordan M. Eizenga
- , Susan Fairley
- , Olivier Fedrigo
- , Adam L. Felsenfeld
- , Xiaowen Feng
- , Christian Fischer
- , Paul Flicek
- , Giulio Formenti
- , Adam Frankish
- , Robert S. Fulton
- , Yan Gao
- , Shilpa Garg
- , Erik Garrison
- , Nanibaa' A. Garrison
- , Carlos Garcia Giron
- , Richard E. Green
- , Cristian Groza
- , Andrea Guarracino
- , Leanne Haggerty
- , Ira M. Hall
- , William T. Harvey
- , Marina Haukness
- , David Haussler
- , Simon Heumos

- , Glenn Hickey
- , Kendra Hoekzema
- , Thibaut Hourlier
- , Kerstin Howe
- , Miten Jain
- , Erich D. Jarvis
- , Hanlee P. Ji
- , Eimear E. Kenny
- , Barbara A. Koenig
- , Alexey Kolesnikov
- , Jan O. Korbel
- , Jennifer Kordosky
- , Sergey Koren
- , HoJoon Lee
- , Alexandra P. Lewis
- , Heng Li
- , Wen-Wei Liao
- , Shuangjia Lu
- , Tsung-Yu Lu
- , Julian K. Lucas
- , Hugo Magalhães
- , Santiago Marco-Sola
- , Pierre Marijon
- , Charles Markello
- , Tobias Marschall
- , Fergal J. Martin
- , Ann McCartney
- , Jennifer McDaniel
- , Karen H. Miga
- , Matthew W. Mitchell
- , Jean Monlong
- , Jacquelyn Mountcastle
- , Katherine M. Munson
- , Moses Njagi Mwaniki
- , Maria Nattestad
- , Adam M. Novak
- , Sergey Nurk

- , Hugh E. Olsen
- , Nathan D. Olson
- , Benedict Paten
- , Trevor Pesout
- , Adam M. Phillippy
- , Alice B. Popejoy
- , David Porubsky
- , Pjotr Prins
- , Daniela Puiu
- , Mikko Rautiainen
- , Allison A. Regier
- , Arang Rhie
- , Samuel Sacco
- , Ashley D. Sanders
- , Valerie A. Schneider
- , Baergen I. Schultz
- , Kishwar Shafin
- , Jonas A. Sibbesen
- , Jouni Sirén
- , Michael W. Smith
- , Heidi J. Sofia
- , Ahmad N. Abou Tayoun
- , Françoise Thibaud-Nissen
- , Chad Tomlinson
- , Francesca Floriana Tricomi
- , Flavia Villani
- , Mitchell R. Vollger
- , Justin Wagner
- , Brian Walenz
- , Ting Wang
- , Jonathan M. D. Wood
- , Aleksey V. Zimin
- & Justin M. Zook

## Contributions

Conceptualization and design: M.R.V., K. Harris, W.S.D., P.H. and E.E.E. Identification and analysis of SNVs from phased assemblies: M.R.V. Mutational spectrum analysis: M.R.V., W.S.D., M.E.G. and K. Harris. Evolutionary age analysis: M.R.V. and P.H. Assembly generation: M.A., J.L., B.P. and HPRC. PacBio genome sequence generation: K.M.M., A.P.L., K. Hoekzema and G.A.L. Copy number analysis and validation: P.C.D., X.G., W.T.H., A.N.R., D. Porubsky and M.R.V. Table organization: M.R.V. Supplementary material organization: M.R.V. Display items: M.R.V., X.G., P.H. and P.C.D. Resources: HPRC, K. Harris, B.P. and E.E.E. Manuscript writing: M.R.V. and E.E.E. with input from all authors.

### Corresponding author

Correspondence to Evan E. Eichler.

## Ethics declarations

### Competing interests

E.E.E. is a scientific advisory board member of Variant Bio, Inc. All other authors declare no competing interests.

## Peer review

### Peer review information

*Nature* thanks Anna Lindstrand and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

## Additional information

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Extended data figures and tables

## Extended Data Fig. 1 Analysis schema for variant and IGC calling.

Whole-genome alignments were calculated for the HPRC assemblies against T2T-CHM13 v1.1 with a copy of GRCh38 chrY using minimap2 v2.24. The alignments were further processed to remove alignments that were redundant in query sequence or that had structural variants over 10 kbp in length. After these steps, the remaining alignments over 1 Mbp were defined to be syntenic and used in downstream analyses. We then counted all pairwise single-nucleotide differences between the haplotypes and the reference and stratified these results into unique regions versus SD regions based on the SD annotations from T2T-CHM13 v1.1. All variants intersecting tandem repeats were filtered to avoid spurious SNV calls. To detect candidate regions of IGC, the query sequence with syntenic alignments was fragmented into 1 kbp windows with a 100 bp slide and realigned back to T2T-CHM13 v1.1 independent of the flanking sequence using minimap2 v2.24 to identify each window's single best alignment position. These alignments were compared to their original syntenic alignment positions, and if they were not overlapping, we considered them to be candidate IGC windows. Candidate IGC windows were then merged into larger intervals and realigned when windows were overlapping in both the donor and the acceptor sequence. We then used the CIGAR string to identify the number of matching and mismatching bases at the "donor" site and compared that to the number of matching and mismatching bases at the acceptor site determined by the syntenic alignment to calculate the number of supporting SNVs.

## Extended Data Fig. 2 Ideogram of an assembly of CHM1 aligned to T2T-CHM13.

The ideogram depicts the contiguity (alternating blue and orange contigs) of a CHM1 assembly generated by Verkko as compared to T2T-CHM13. The overall contig N50 is 105.2 Mbp providing near chromosome arm contiguity with the exception of breaks at the centromere (red) and other large satellite arrays. Because the sequence is derived from a monoploid complete hydatidiform mole, there is no opportunity for assembly errors due to inadvertent haplotype switching.

## Extended Data Fig. 3 Increased variation in SD sequences and African haplotypes.

Histograms of the average number of SNVs per 10 kbp over all 125 Mbp bins of unique (blue) and SD (red) sequence for all haplotypes. African haplotypes (bottom) are compared separately to non-African (top) haplotypes. All SD bins (125 Mbp each) have more SNVs than any unique bin irrespective of human superpopulation.

## Extended Data Fig. 4 Average number of SNVs across different repeat classes.

Shown are the average number of SNVs per 10 kbp within SDs (red), unique (blue), and additional sequence classes (gray) across the HPRC haplotypes. These classes include exonic regions, ancient SDs (SD with <90% sequence identity) and all elements identified by RepeatMasker (RM) with Alu, L1 LINE, and HERV elements broken out separately. Below each sequence class we show the average number of SNVs per 10 kbp for the median haplotype. Standard deviations and measurements for additional repeat classes are provided in Table S3.

## Extended Data Fig. 5 Largest IGC events in the human genome.

The ideogram depicts as red arcs the positions of the largest IGC events between and within human chromosomes (top 10% of the length distribution).

## Extended Data Fig. 6 Percent of increased single-nucleotide variation explained by IGC.

Shown is the fraction of the increased SNV diversity in SDs that can be attributed to IGC for each of the HPRC haplotypes stratified by global superpopulation. In text is the average across all haplotypes (23%).

## Extended Data Fig. 7 IGC hotspots.

**a)** Density of IGC acceptor (top, blue) and donor (bottom, orange) sites across the "SD genome". The SD genome consists of all main SD regions (>50 kbp) minus the intervening unique sequences. **b)** All intrachromosomal IGC events from 102 human haplotypes analyzed for chromosome 15. Arcs drawn in blue (top) have the acceptor site on the left-hand side and the donor site on the right. Arcs drawn in orange (bottom) are arranged oppositely. Protein-coding genes are drawn as vertical black lines above the ideogram, and large duplication (blue) and deletion (red) events associated with human diseases are drawn as horizontal lines just above the ideogram. **c)** Zoom of the 100 highest confidence (lowest p-value) IGC events identified on chromosome 15 between 17 and 31 Mbp. Genes that are intersected by IGC events are highlighted in red.

## Supplementary information

### Supplementary Information

This file contains Supplementary Figs. 1–20, Notes and References.

### Reporting Summary

### Supplementary Tables

This file contains Supplementary Tables 1–14.

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or

exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

# About this article

## Cite this article

Vollger, M.R., Dishuck, P.C., Harvey, W.T. *et al.* Increased mutation and gene conversion within human segmental duplications. *Nature* **617**, 325–334 (2023). https://doi.org/10.1038/s41586-023-05895-y

- Received: 06 July 2022
- Accepted: 28 February 2023
- Published: 10 May 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05895-y

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

## Human pangenome supports analysis of complex genomic regions

- Arya Massarat
- Melissa Gymrek
- Hákon Jónsson

Nature News & Views Forum 10 May 2023

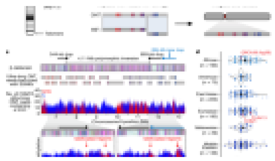
## A draft human pangenome reference

- Wen-Wei Liao
- Mobin Asri
- Benedict Paten

Nature Article Open Access 10 May 2023

---

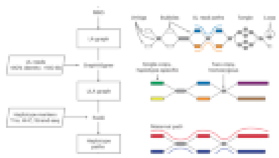
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05895-y



## The structure, function and evolution of a complete human chromosome 8

07 April 2021

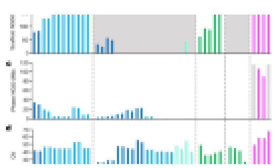
Glennis A. Logsdon, Mitchell R. Vollger, … Evan E. Eichler



## Telomere-to-telomere assembly of diploid chromosomes with Verkko

16 February 2023

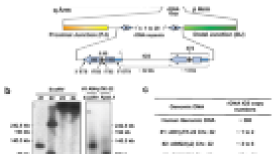
Mikko Rautiainen, Sergey Nurk, … Sergey Koren



## Semi-automated assembly of high-quality diploid human reference genomes

19 October 2022

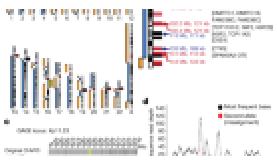
Erich D. Jarvis, Giulio Formenti, … Human Pangenome Reference Consortium



## [The genomic structure of a human chromosome 22 nucleolar organizer region determined by TAR cloning]

04 February 2021

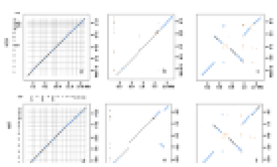
Jung-Hyun Kim, Vladimir N. Noskov, … Vladimir Larionov



## [Telomere-to-telomere assembly of a complete human X chromosome]

14 July 2020

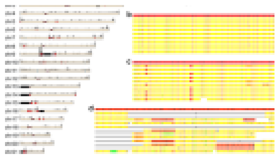
Karen H. Miga, Sergey Koren, … Adam M. Phillippy



## [Telomere-to-telomere assembled and centromere annotated genomes of the two main subspecies of the button mushroom Agaricus bisporus reveal especially polymorphic chromosome ends]

04 September 2020

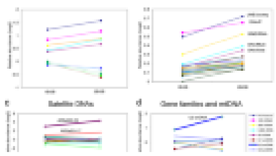
Anton S. M. Sonnenberg, Narges Sedaghat-Telgerd, … A. van Peer



## [Genome maps across 26 human populations reveal population-specific patterns of structural variation]

04 March 2019

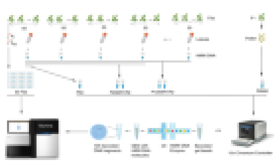
Michal Levy-Sakin, Steven Pastor, … Pui-Yan Kwok



## [Out of patterns, the euchromatic B chromosome of the grasshopper Abracris flavolineata is not enriched in high-copy repeats]

04 September 2021

Diogo Milani, Francisco J. Ruiz-Ruano, … Diogo C. Cabral-de-Mello



## [Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination]

20 September 2019

Hequan Sun, Beth A. Rowan, … Korbinian Schneeberger

- Article
- Open Access

- Published: 10 May 2023

# Recombination between heterologous human acrocentric chromosomes

- Andrea Guarracino,
- Silvia Buonaiuto,
- Leonardo Gomes de Lima,
- Tamara Potapova,
- Arang Rhie,
- Sergey Koren,
- Boris Rubinstein,
- Christian Fischer,
- Human Pangenome Reference Consortium,
- Jennifer L. Gerton,
- Adam M. Phillippy,
- Vincenza Colonna &
- …
- Erik Garrison

Show authors

*Nature* **volume 617**, pages 335–343 (2023)

- 294 Altmetric

- Metrics details

## Abstract

The short arms of the human acrocentric chromosomes 13, 14, 15, 21 and 22 (SAACs) share large homologous regions, including ribosomal DNA repeats and extended segmental duplications[1,2]. Although the resolution of

these regions in the first complete assembly of a human genome—the Telomere-to-Telomere Consortium's CHM13 assembly (T2T-CHM13)—provided a model of their homology[3], it remained unclear whether these patterns were ancestral or maintained by ongoing recombination exchange. Here we show that acrocentric chromosomes contain pseudo-homologous regions (PHRs) indicative of recombination between non-homologous sequences. Utilizing an all-to-all comparison of the human pangenome from the Human Pangenome Reference Consortium[4] (HPRC), we find that contigs from all of the SAACs form a community. A variation graph[5] constructed from centromere-spanning acrocentric contigs indicates the presence of regions in which most contigs appear nearly identical between heterologous acrocentric chromosomes in T2T-CHM13. Except on chromosome 15, we observe faster decay of linkage disequilibrium in the pseudo-homologous regions than in the corresponding short and long arms, indicating higher rates of recombination[6,7]. The pseudo-homologous regions include sequences that have previously been shown to lie at the breakpoint of Robertsonian translocations[8], and their arrangement is compatible with crossover in inverted duplications on chromosomes 13, 14 and 21. The ubiquity of signals of recombination between heterologous acrocentric chromosomes seen in the HPRC draft pangenome suggests that these shared sequences form the basis for recurrent Robertsonian translocations, providing sequence and population-based confirmation of hypotheses first developed from cytogenetic studies 50 years ago[9].

# Main

Although the human reference genome is now 22 years old[10], fundamental limitations of the bacterial artificial chromosome (BAC) libraries on which it was built prevented its completion. Incomplete regions amount to 8% of the Genome Reference Consortium's Human Build 38, and include heterochromatic regions in the centromeres and the SAACs. Advances in long-read DNA sequencing have recently enabled the creation of a complete reference assembly—T2T-CHM13[3]—from a homozygous human cell line, providing a reference system for these regions for the first time. In parallel, our ongoing work in the HPRC has yielded 94 haplotype-resolved assemblies for human cell lines (HPRCy1) based on the same Pacific

Biosciences circular consensus (HiFi) sequencing that forms the foundation of T2T-CHM13[4]. These resources enable us to characterize patterns of variation in these previously invisible regions. Here we study variation in the largest non-centromeric regions made visible in T2T-CHM13 and HPRCy1—those between the centromere and the ribosomal DNA (rDNA) on the SAACs, where Robertsonian translocations[11] (ROBs), the most common human translocation events, frequently occur.

Eighteen of the twenty-three human chromosomes are metacentric, with the centromere found in a median position between short (p) and long (q) arms, whereas 5 are acrocentric, featuring one arm that is substantially shorter than the other. The SAACs (chromosome (chr.)13p, chr. 14p, chr. 15p, chr. 21p and chr. 22p) host the nucleolus organizer regions, the genomic segments that contain rDNA genes and that give rise to the interphase nucleoli[12,13]. Owing to their repetitive nature, rDNA repeat arrays facilitate intramolecular recombination[14]. rDNA repeats incur double-strand breaks at a high rate owing to transcription–replication conflicts[13]. Moreover, rDNA from multiple acrocentric chromosomes can be co-located in nucleoli during interphase, and multiple acrocentric chromosomes often co-localize to a single nucleolus during the pachytene stage of meiosis[15], when chromosomes synapse and recombine. As it causes them to occupy the same constrained physical space, the positioning of rDNA-adjacent sequences in proximity to the nucleolus could be a driver of genetic exchange between heterologous chromosomes (Supplementary Note 1), given that estimates[15] of the probability of two regions adjacent to the nucleolus-organizing region being colocalized are 120,000 times higher than colocalization in a human spermatocyte nucleus. In line with this, distal and proximal sequences to rDNA repeat arrays are conserved among the acrocentric chromosomes, suggesting that recombination homogenizes them[1,2]. Experimental and sequence-based evidence indicates the presence of a common subfamily of alpha satellite DNA shared by acrocentric pairs chr. 13–chr. 21 and chr. 14–chr. 22 that provides evidence for an evolutionary process consistent with recombination between heterologous chromosomes[16]. Furthermore, ROBs—which occur in 1 out of 800 births—are most common between chr. 13 and chr. 14 (around 75% of cases), and between chr. 14 and chr. 21 (around 10% of cases), but the underlying sequences and recombination processes that drive them remain unknown[8].

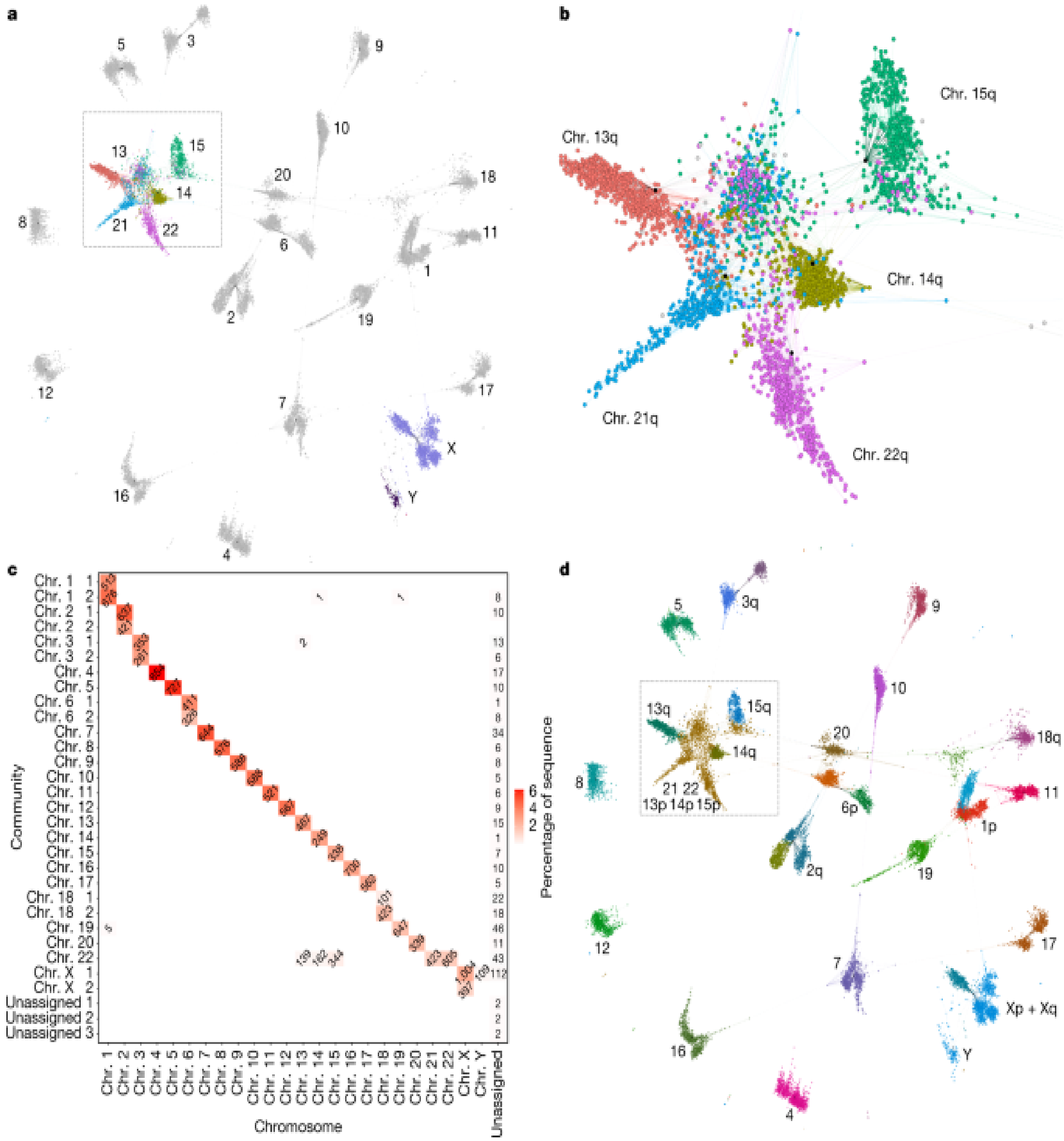
The T2T-CHM13 reference fully resolves the genomic structure of the SAACs, confirming their strong similarity and providing a complete view of the homologies in this single genome[3]. However, T2T-CHM13 does not provide information on how SAACs vary among the human population and additional genomes are needed to understand whether the representation in T2T-CHM13 is typical. Notably, alignments of HPRCy1 assemblies to T2T-CHM13 reveal individual contigs with optimal alignments to multiple CHM13 acrocentric chromosomes, suggesting possible translocations[4]. This analysis was necessarily relative to only a single frame of reference used as target in alignment, leaving open questions regarding the relationships between pairs of HPRCy1 haplotypes. A complete study of this region thus requires improvements in both sequence assembly and pangenome analysis to enable an unbiased assessment of its structure and variation in the population. Here we combine T2T-CHM13 and HPRCy1 assemblies in a reference-free pangenome variation graph (PVG) model of the SAACs. Using this model and other symmetric analyses of T2T-CHM13 and the HPRCy1 assemblies, we establish a coherent model of population-scale variation in the SAACs.

# Chromosome community detection

We sought to study the chromosome groupings implied by the homologies found in all 94 assemblies of the HPRCy1 pangenome. We used homology mapping to build a reference-free model of the HPRCy1 pangenome, represented as a mapping graph with nodes as contigs and edges as mappings between them. The graph was built using chains of 50-kb seeds of 95% average nucleotide identity—features that we expect to support homologous recombination[17]—with up to 93 alternative mappings allowed per contig. After applying this process to all 38,325 HPRCy1 contigs and narrowing our focus to only mappings involving contigs that are at least 1 Mb in size, we built a reduced mapping graph by selecting the best 3 mappings per contig segment and labelling each contig with its reference-relative assignment (Fig. 1a). This simplified graph showed clusters that generally matched our expectations of higher similarity between certain chromosomes[18,19] (Fig. 1b). For a more quantitatively rigorous interpretation, we used a community detection algorithm (Methods) to divide the full mapping graph into 31 communities (Supplementary File 1).

These communities were consistent with our expectations based on mapping the contigs to reference chromosomes T2T-CHM13 and GRCh38 and known patterns of similarity between chromosomes. We found that the community of the SAACs contained the most distinct chromosomes and the most contigs (Fig. 1c,d). Many contigs from the pseudoautosomal regions (PARs) and X-transposed regions (XTRs) of chromosome X[20] and all of those from chromosome Y formed one community, and others from the short arm of chromosome X—including all of those from evolutionary strata 4 and 5[21]—formed another (Extended Data Fig. 1). A few additional communities were identified that did not correspond to individual chromosomes, but typically represent single chromosome arms.

**Fig. 1: Community detection in the HPRCy1 pangenome.**




**a**, The reduced all-to-all mapping graph of HPRCy1 against itself, with contigs represented as nodes and mappings as edges. Colours distinguish the acrocentric or sex chromosome to which each contig was assigned by competitive mapping against T2T-CHM13 and GRCh38, with text labels indicating the chromosome for each visual cluster. **b**, A close-up view of the region indicated in **a** and **d** containing nearly all contigs that match acrocentric chromosomes. **c**, Results of community assignment on the mapping graph. The *x*-axis shows the chromosome to which contigs belong, based on competitive mapping to T2T-CHM13 and GRCh38; the *y*-axis

indicates the community, which is named according to the chromosome that contributes the largest number of contigs to it. In the squares, the numbers indicate how many contigs belong each specific chromosome and community and the shade indicates the percentage of the total assembly sequence present in the set. The sex chromosomes and the acrocentric chromosomes participate in the only clusters that mix many (more than 100) contigs belonging to different chromosomes. **d**, The reduced homology mapping graph in **a**, coloured according to community assignment (colours do not correlate with those in **a** or **b**). The p-arms of chr. 13, chr. 14 and chr. 15, and all of chr. 21 and chr. 22 form one community, and chr. Y and most of chr. X form another.

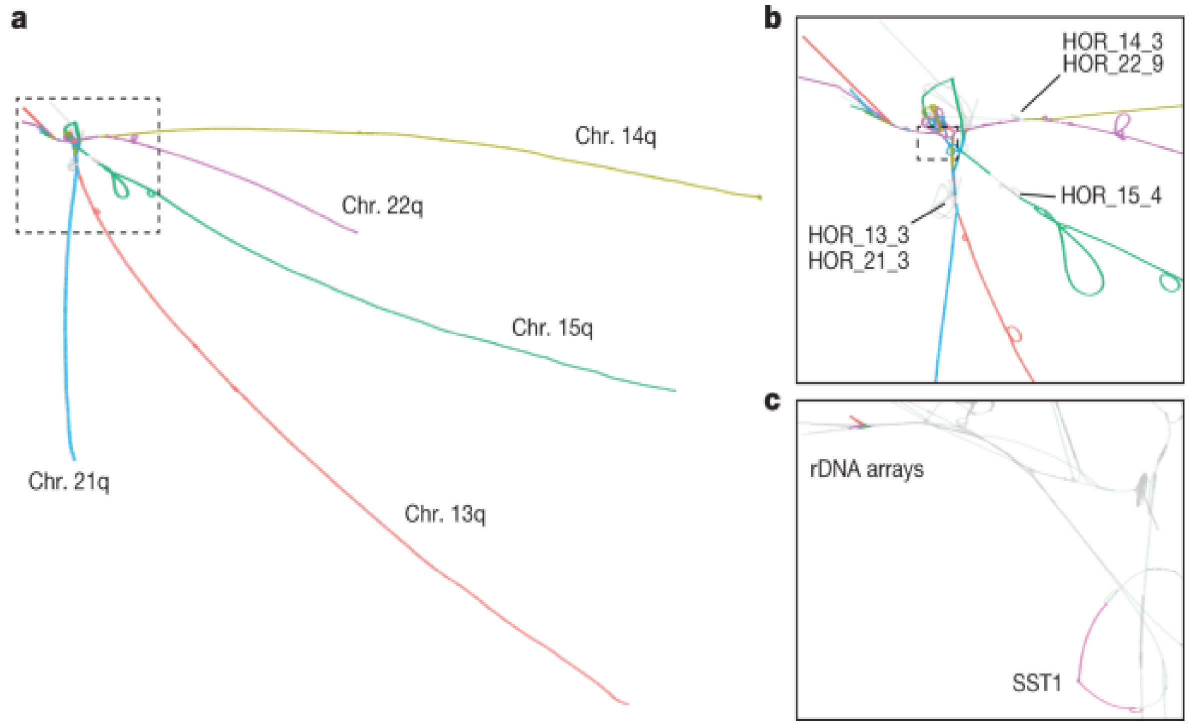
# An all-acrocentric PVG

We constructed a pangenome graph from acrocentric contigs in the HPRCy1 draft pangenome to evaluate the hypothesis that heterologous SAACs recombine. We first collected long HPRCy1 contigs that span the acrocentric centromeres and can be assigned to specific acrocentric chromosomes (Extended Data Fig. 2). We then used the PanGenome Graph Builder[22] (PGGB) to construct a single PVG from these contigs (Methods). PVG nodes represent sequences and edges indicate when concatenations of the nodes they connect occur in the contigs represented by the graph[23]. By relating pangenome sequences to the graph as paths of nodes[5], PVGs support base-level analysis of variation and homology between genomes[4,24,25,26]. The symmetric all-to-all alignment[27] and graph induction[28] of PGGB avoid sources of bias such as reference choice and genome inclusion order that affect progressive PVG construction methods[28]. For cross-validation of our results, we additionally include two assemblies of HG002 in the PVG: HG002-HPRCy1[4]—obtained from HiFi reads, and HG002-Verkko—a T2T diploid assembly constructed from both HiFi and Oxford Nanopore Technologies (ONT) reads as described in Methods.

The resulting acrocentric PVG (acro-PVG) presents structures that echo those observed in T2T-CHM13 and the community structure of the homology mapping graph (Fig. 2a and Supplementary Files 2 and 3). In more detail, the main connected component including all chromosomes

presented a tangled region, anchored at the rDNA repeats and extending towards the centromere-proximal end of the short arms. The alpha satellite higher-order repeat arrays in the centromeres of chr. 13–chr. 21 and chr. 14–chr. 22 pairs shared high similarity within each pair[18,19], leading to collapsed motifs in the graph (Fig. 2b). The chr. 13–chr. 21 and chr. 14–chr. 22 pairs diverge in centromere-proximal regions of the q-arms. Furthermore, a region in the pangenome graph centred on the GC-rich SST1 array was present in a single copy in chr. 13, chr. 14 and chr. 21, indicating a high degree of similarity of genomes in those regions (Fig. 2c and Supplementary Fig. 1). This is compatible with the frequent involvement of these regions of chr. 13, chr. 14 and chr. 21 in ROBs[8,29]. The SST1 elements in the segmentally duplicated region are GC-rich 1.4- to 2.4-kb-long sequences arranged in tandem clusters[30], located throughout the genome including near the centromeres of the SAACs chr. 13, chr. 14 and chr. 21[31]. The SST1 array size is variable in the human population[32] and its methylation status is clinically relevant to cancer[33]. SST1 repeats on chr. 13, chr. 14 and chr. 21 in T2T-CHM13 are highly similar to each other[31], consistent with homogenization via recombination. All the graph motifs described in the acro-PVG were also confirmed by building a pangenome graph without including the T2T-CHM13 and GRCh38 references (Supplementary Figs. 2 and 3), indicating that the observed structure is independent of the reference assemblies.

**Fig. 2: The acro-PVG derived from the HPRCy1 assembly.**




**a**, The major component of the acro-PVG, shown with nodes in T2T-CHM13 chromosomes labelled with the colour scheme from Fig. 1a. The acrocentric q-arms are almost completely separated, whereas the p-arms unite in a tangle adjacent to the rDNA array. **b**, A close-up view of the SAAC junction, showing the separation of centromeric high-order repeats of chr. 15 (HOR_15_4) from the other chromosomes, whereas chr. 13 and chr. 21, and chr. 14 and chr. 22 share substantial homology in their arrays, which causes them to collapse in the PVG. A few assemblies span the rDNA array into its distal junction, which presents as a single homologous region across all chromosomes[2], and then fray into diverse sequences visible as tips in the top left. **c**, Closer view of the outlined region in **b**, focusing on the segmentally duplicated core centred in the SST1 array and the rDNA arrays, as labelled in T2T-CHM13. The highlighted region around the SST1 array is in the same orientation on T2T-CHM13 chr. 13p11.2 and chr. 21p11.2, and is inverted on chr. 14p11.2; these 3 regions have a pairwise identity[3] of more than 99%.

## Exchange among heterologous acrocentric regions

The acro-PVG provides a representation of the multiple alignment of SAACs found in the human population. In the acro-PVG, we observe many regions in the graph where multiple T2T-CHM13 chromosomes are aligned. We expect these regions to potentially support homologous recombination, which largely depends on sequence homology and physical proximity, both of which are common among heterologous SAACs[15,34].

## HPRCy1 contigs are homology mosaics

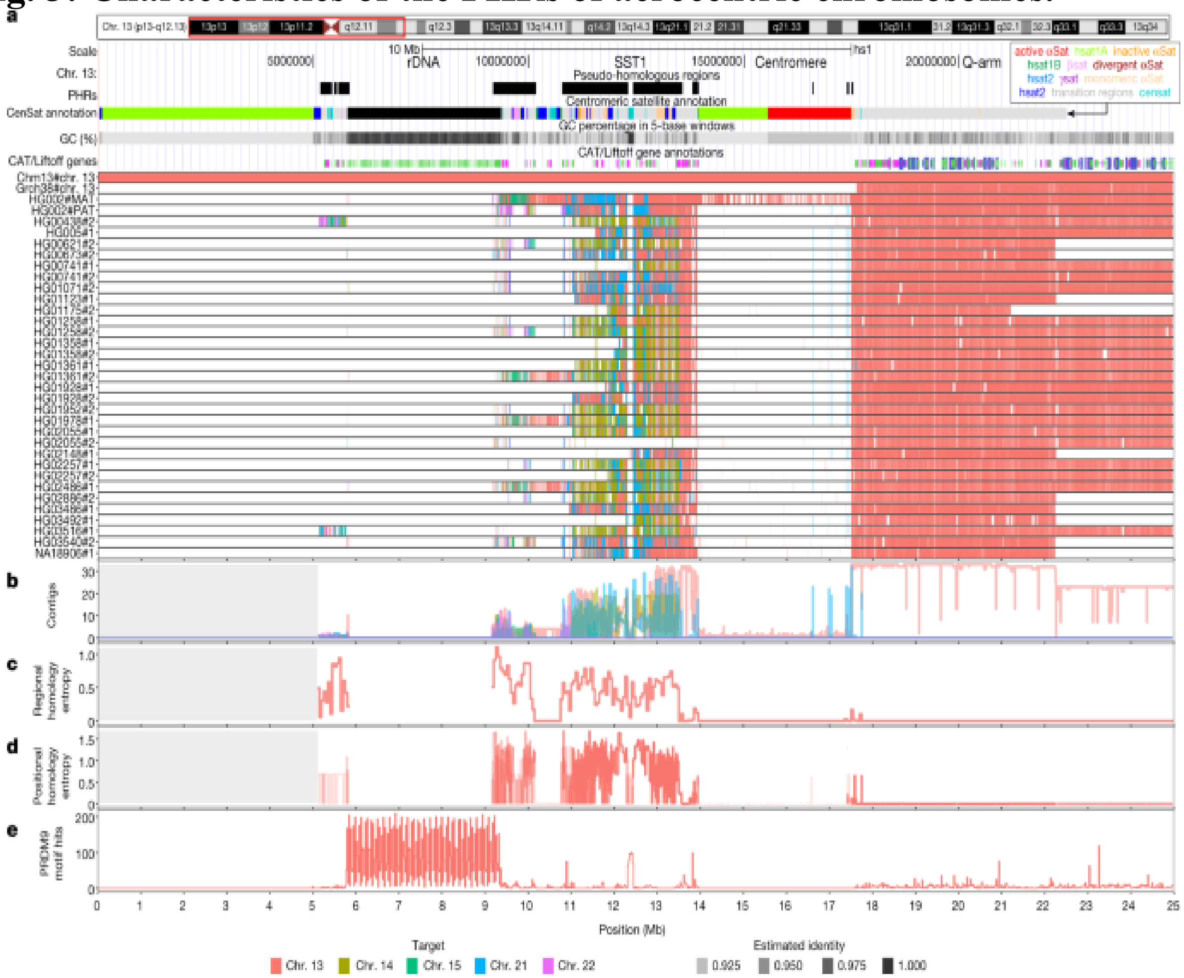
We sought to test the hypothesis that homologous regions of the SAACs feature ongoing sequence exchange by searching for regions in the acro-PVG where individual contigs are best described as a mosaic of diverse T2T-CHM13 acrocentric chromosomes. We derived a pairwise alignment from the acro-PVG through 'untangling'[26], a process that projects the graph into an alignment between a set of query (HPRCy1 acrocentric (HPRCy1-acro)) and reference (T2T-CHM13) sequences, jointly considering all possible alignments represented by the pangenome graph. The untangling of the acro-PVG against multiple T2T-CHM13 chromosome reference sequences simultaneously shows the best match of segments within contigs to multiple reference chromosomes.

The hypothesis of recombination between heterologous acrocentric chromosomes implies that the HPRCy1-acro contigs untangled from the acro-PVG will be a mosaic of diverse acrocentric chromosomes in the regions undergoing homologous recombination. The same would not be true for flanking regions that should map to one specific chromosome.

We queried the PVG[26] to obtain a mapping from segments of all PVG paths onto T2T-CHM13. This segments the graph, and for each HPRCy1 contig (query) subpath through each graph segment, we find the most similar reference segment (Extended Data Fig. 3). To reduce the possibility of error, we focused the alignment projection only on the confidently assembled regions of the HPRCy1-acro contigs[4] (Methods) and we filtered the mappings to retain only those at greater than 90% estimated identity, removing a total of 1.17 Gbp, or 2.52% of the total SAAC contig segments (Supplementary File 4 and Supplementary Figs. 4–8).

For a reference-relative interpretation of the results, we anchored the contigs to the single T2T-CHM13 reference chromosomes to which the q-arm maps (Methods), providing a reference-relative positioning of contigs in the PVG. We find that the q-arm of each contig maps to a single chromosome, whereas the p-arm is a mosaic of segments mapping to several acrocentric chromosomes (Fig. 3a,b). Results for all the acrocentric chromosomes are shown in Extended Data Figs. 4–8.

**Fig. 3: Characteristics of the PHRs of acrocentric chromosomes.**




**a**, We focus on the first 25 Mb of chr. 13, shown here as a red box over T2T-CHM13 cytobands. PHRs are highlighted relative to T2T-CHM13 genome annotations for centromere and satellite repeats (CenSat annotation), GC percentage and genes (CAT/Liftoff genes). Top, regions of interest described in the main text: rDNA, the SST1 array, the centromere and q-arm. Bottom, relative homology mosaics based on the T2T-CHM13 assembly for each chr. 13-matched contig from HPRCy1-acro, with colours indicating the most

similar reference chromosome (target). **b**,**c**,**d**, Aggregated untangle results in the SAACs. **b**, The count of HPRCy1 q-arm-anchored contigs mapping to each acrocentric chromosome (Contigs) aggregated by target chromosome and (**c**) the regional (50 kb) untangle entropy metric (Regional homology entropy) computed over the contigs' untangling relative to T2T-CHM13. **d**, By considering the multiple untangling of each HPRCy1-acro contig, we develop a point-wise metric that captures diversity in homology patterns relative to T2T-CHM13 (Positional homology entropy), leading to our definition of the PHRs. **e**, The patterns of homology mosaicism suggest ongoing recombination exchange in the SAACs. A scan over T2T-CHM13 reveals that the rDNA and SST1 array units are enriched for PRDM9 binding motifs, and thus may host frequent double-stranded breaks during meiosis. In **b**–**d**, the grey background indicates regions with missing data due to the lack of non-T2T-CHM13 contigs.

We cross-validated homology mosaic patterns by comparing the reference-relative anchored untangling of HG002-Verkko to HG002-HiFi, obtaining a 87.45% concordance rate in the SAACs and a 99.93% concordance rate in the acrocentric q-arms (Methods). Although HG002-HiFi contains only one contig that would meet our HPRCy1 contiguity requirements, we observe broadly concordant patterns in the two assemblies (Supplementary Figs. 9–13) and visually confirm patterns—such as those between chr. 13p and chr. 21p—that are seen in many HPRCy1 assemblies (Supplementary Fig. 9).

## Homology mosaicism grows across SAACs

We counted the number of contigs that best match each of the T2T-CHM13 acrocentric chromosomes within the PVG (Fig. 3c). On the q-arm, all contigs best match their homologous T2T-CHM13 chromosome, agreeing with the observed structure of the PVG (Fig. 2a). However, as we approach the centromere from the q-arm, we observe regions of homology between chr. 13 and chr. 21 (Fig. 3c and Extended Data Figs. 4b and 7b) and between chr. 14 and chr. 22 (Extended Data Figs. 5b and 8b). By contrast, homology with other acrocentric chromosomes begins closer to the rDNA in chr. 15 (Extended Data Fig. 6b), corroborating the pattern observed in the PVG topology (Fig. 2b).

Although the higher-order repeat arrays on chr. 13 and chr. 21 and on chr. 14 and chr. 22 are both collapsed in the PVG (Fig. 2b), we observe sparse identity mappings higher than 90% within the centromeres (Fig. 3b and Extended Data Figs. 4a, 5a, 6a, 7a and 8a). This is consistent with other reports of high divergence within centromeric satellites[35]. HPRCy1 contigs anchored on the q-arms of chr. 13, chr. 14 and chr. 21 share a segmental duplication (or homologous region) centred on the SST1 array (Fig. 3b), in line with what is seen in the pangenome graph topology (Fig. 2c and Supplementary Figs. 1 and 3). Furthermore, as in T2T-CHM13, this region is in the same orientation on chr. 13 and chr. 21, but is inverted on chr. 14 (Supplementary Figs. 14–16). All chromosomes provide similarly good matches for contigs in the regions immediately proximal and distal to the rDNA. However, this is supported by relatively few (nine) q-arm-anchored contigs that purport to cross the rDNA—loci that we do not expect to assemble correctly using current sequencing approaches[3].

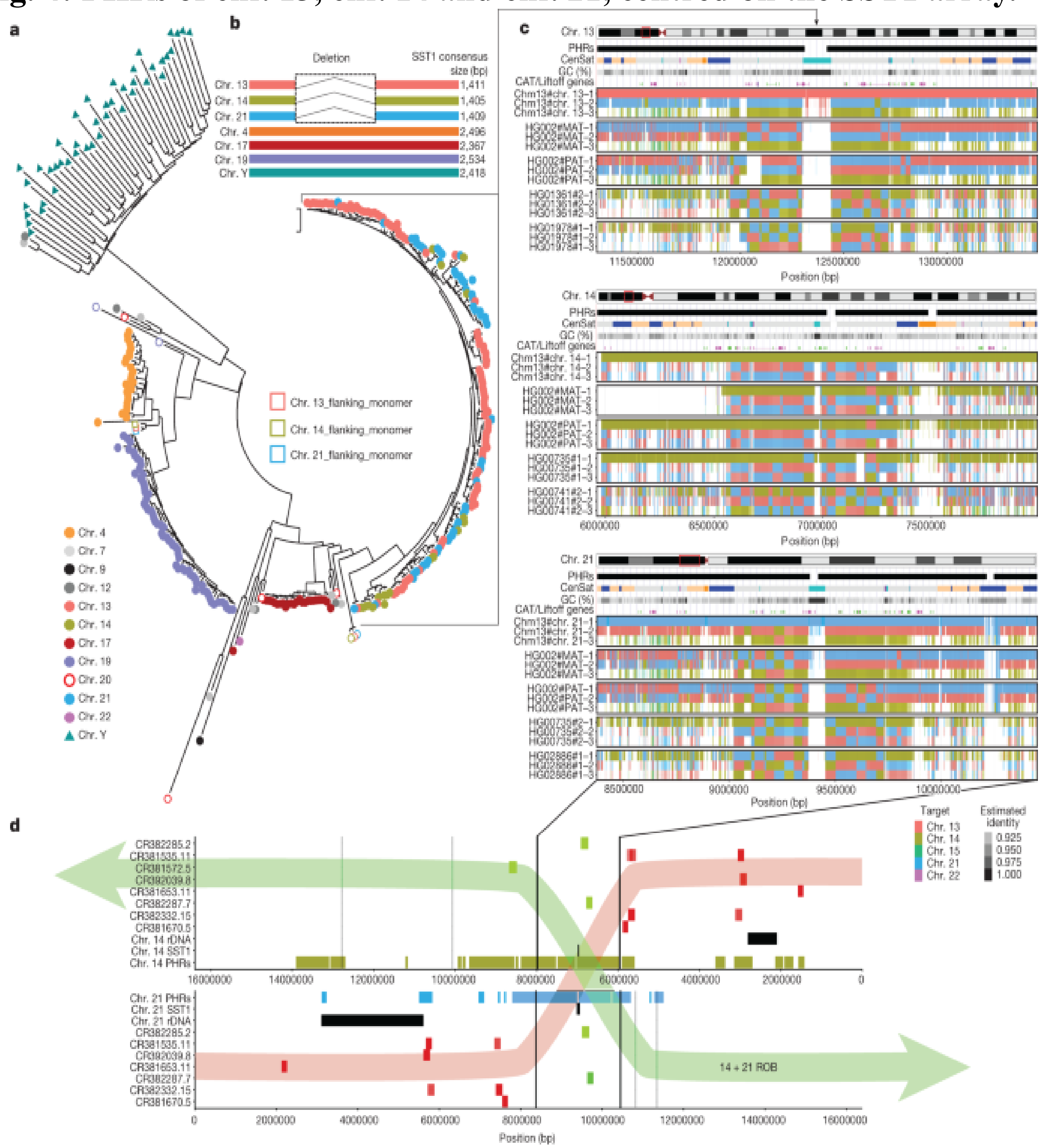
To assess the homology between the acrocentric chromosomes, we developed a metric that captures the degree of disorder in the untangling of HPRCy1-acro contigs over 50-kb regions of T2T-CHM13 (Methods). This metric, ‘regional homology entropy’, is greater than 0 in regions where contigs match multiple T2T-CHM13 chromosomes—a pattern indicative of recombination. We find that regional homology entropy increases as we progress over each short arm and reaches a maximum immediately on the proximal flanks of the rDNA arrays (Fig. 3d). We observed an equivalent increase of regional homology entropy in the PARs on chr. X and chr. Y (Supplementary Fig. 17), which are known to actively recombine.

# Acrocentric PHRs

Our analyses suggest that regions of near-identity between multiple T2T-CHM13 chromosomes are capable of supporting large-scale homologous recombination. To study the boundaries of these regions, we derived a multiple untangling, which orders by identity multiple T2T-CHM13 matches for every contig segment (Supplementary Figs. 18–22). The order of T2T-CHM13 hits captures the leaf order of a HPRCy1 contig-rooted phylogeny[36]. Differences in chromosome-relative phylogenies across haplotypes indicate different evolutionary histories and imply ongoing

recombination[37], which leads to chequerboard patterns in the multiple untangling plots (Fig. 4c). To delineate regions where heterologous chromosomes are likely to recombine, we computed ‘positional homology entropy’—a measure of the diversity of reference-relative phylogenies—for each position in T2T-CHM13 (Fig. 3e and Supplementary Fig. 23). We consider regions with positional homology entropy greater than 0 over more than 30 kb to be candidates for ongoing recombination (Methods).

**Fig. 4: PHRs of chr. 13, chr. 14 and chr. 21, centred on the SST1 array.**



**a**, Maximum likelihood phylogenetic analysis of SST1 full-length elements indicates a recent homogenization process of acrocentric arrays. Coloured circles next to chromosome labels indicate individual monomers retrieved from the T2T-CHM13 assembly. Coloured triangles indicate SST1 full-length monomers retrieved from the HG002 chr. Y assembly. Partial or chimeric monomers flanking chr. 13, chr. 14 and chr. 21 arrays (located around 250 kb from the main array) are labelled as open circles or squares, respectively, coloured according to the corresponding chromosome. **b**, Schematic representation of SST1 consensus alignments, indicating a deletion that is present only in the SST1 unit from arrays on chr. 13, chr. 14 and chr. 21. **c**, Multiple untangling of T2T-CHM13, HG002-Verkko haplotypes and HPRCy1-acro contigs versus T2T-CHM13. Three of the five acrocentric chromosomes are represented. The degree of transparency indicates the estimated identity of the mappings. All mappings above 90% estimated pairwise identity are shown. To enable the display of simultaneous hits to all acrocentric regions, each grouping shows the first three best alternative mappings. SST1 arrays described in **a** are at the centre of a PHR that displays chequerboard patterns indicative of recombination between heterologous acrocentric chromosomes (black arrows link the SST1 arrays in all panels). These patterns are less common on chr. 15 and chr. 22 (Supplementary Figs. 20 and 22). **d**, The PHRs on T2T-CHM13 (yellow and light blue) in relation to BACs localized cytogenetically[8] to recurrent chr. 14–chr. 21 ROB breakpoints. BACs shown in green are found in dicentric Robertsonian chromosomes, whereas those in red are not. Chr. 14 is shown in an inverted orientation aligned to chr. 21 at the breakpoint region suggested experimentally[8]. In a transparent overlay, we propose a retained dicentric chromosome (14+21 ROB, green) and lost (red) products of the studied recurrent translocations.

These PHRs total 18.329 Mb in length (Supplementary File 5) and differ in size by chromosome: chr. 13, 4.53 Mb (Fig. 3b); chr. 14, 6.48 Mb (Extended Data Fig. 5a); chr. 15, 719.25 kb (Extended Data Fig. 6a); chr. 21, 3.79 Mb (Extended Data Fig. 7a); and chr. 22, 2.81 Mb (Extended Data Fig. 8a). We term them PHRs by analogy to the PARs of sex chromosomes, because these homology domains could enable non-homologous chromosomes to pair like homologous chromosomes. Notably, the chromosomes involved in the most common ROBs (chr. 13–chr. 14 and chr. 14–chr. 21) have larger PHRs,

which could promote the recombination events that lead to these translocations. Supporting this, BAC clones surrounding common recurrent ROB breakpoints[8] map to T2T-CHM13 PHRs (Supplementary Fig. 24). A genome-wide phylogenetic analysis of SST1 array elements indicates the expected pattern of concerted evolution by chromosome, but the repeats from chr. 13, chr. 14 and chr. 21 display a unique pattern of concerted evolution between chromosomes (Fig. 4a) and furthermore, share a deletion of around 1.0 kb relative to all other SST1 repeats (Fig. 4b and Supplementary Fig. 25), suggesting inter-array recombination similar to the surrounding non-satellite sequences of the PHRs (Fig. 4c). We confirmed that patterns observed by fluorescent in situ hybridization of these BACs are compatible with breakpoints occurring in the PHRs centred at the SST1 array (Fig. 4d).

To provide a positive control, we applied the same method to the sex chromosome PVG to identify their PHRs (chr. X, 2.75 Mb and chr. Y, 2.73 Mb; Supplementary File 6 and Supplementary Fig. 26). These regions precisely match the established boundaries for the PARs and contain sparse hits in the XTRs, which would be compatible with reports of X–Y interchange in the XTR[38]. Biallelic SNP calls from a whole-genome HPRCy1 graph released in the accompanying Article[4] show that variant density in the PARs and the acrocentric p-arms is markedly higher than elsewhere in these chromosomes (Supplementary Fig. 27), which is consistent with increased rates of recombination in these regions[20].

In humans and many other mammals, the sequence specificities of the DNA-binding zinc finger protein PRDM9 regulate the formation of double-stranded breaks that drive meiotic homologue synapsis and recombination[39,40]. We scanned T2T-CHM13 acrocentric chromosomes for PRDM9 motifs detected by chromatin immunoprecipitation with sequencing[41] (Extended Data Fig. 9 and Supplementary File 7), finding that both rDNA and SST1 arrays are enriched for PRDM9 motifs relative to the surrounding sequence (Fig. 3f and Extended Data Figs. 4e, 5e, 6e, 7e and 8e). By contrast, we find almost no PRDM9 motifs in the centromeres, where meiotic recombination is harmful and suppressed by diverse mechanisms[42].

## Linkage disequilibrium decay in PHRs

To quantify the magnitude of putative recombination in the PHRs, we calculated the rate of the linkage disequilibrium decay between SNPs detected in the acro-PVG[43] (Supplementary Fig. 28). For each acrocentric chromosome, we plot the $R^2$ allele correlation versus distance for three sets of pairs of variants separated by up to 4 kb: variants on the q-arm, on the p-arm, and within the PHRs. The overall trend of linkage disequilibrium decay on the q-arms is similar to trends seen in other datasets that evaluate linkage disequilibrium in humans[44,45]. On chr. 13 the decay of linkage disequilibrium in PHRs and the p-arm is similar and faster than on the q-arm, and on chr. 14, chr. 15 and chr. 22, linkage disequilibrium decay in PHR is even faster on PHRs compared with the p-arm. The same trend does not apply to chr. 21 (Extended Data Fig. 10).

The fast linkage disequilibrium decay in the PHRs compared to q-arms supports the hypothesis of ongoing recombination exchange. In general, there is a higher level of linkage disequilibrium on the p-arms than on the q-arms, perhaps owing to lower recombination in heterochromatic regions[46]. However, for the majority of acrocentric chromosomes, we observe the opposite in the PHRs. This effect is clearest within the chromosomes that other analyses suggest share the most homologous sequence: chr. 13, chr. 14, chr. 21 and chr. 22, whereas chr. 15—which appears to be an outlier[7] (as also observed in Figs. 1 and 2)—contains shorter PHRs and we have less confidence in the linkage disequilibrium decay trends (error bars on Extended Data Fig. 10, chr. 15 PHRs). This pattern is consistent with an increased recombination rate in the PHRs[6,7].

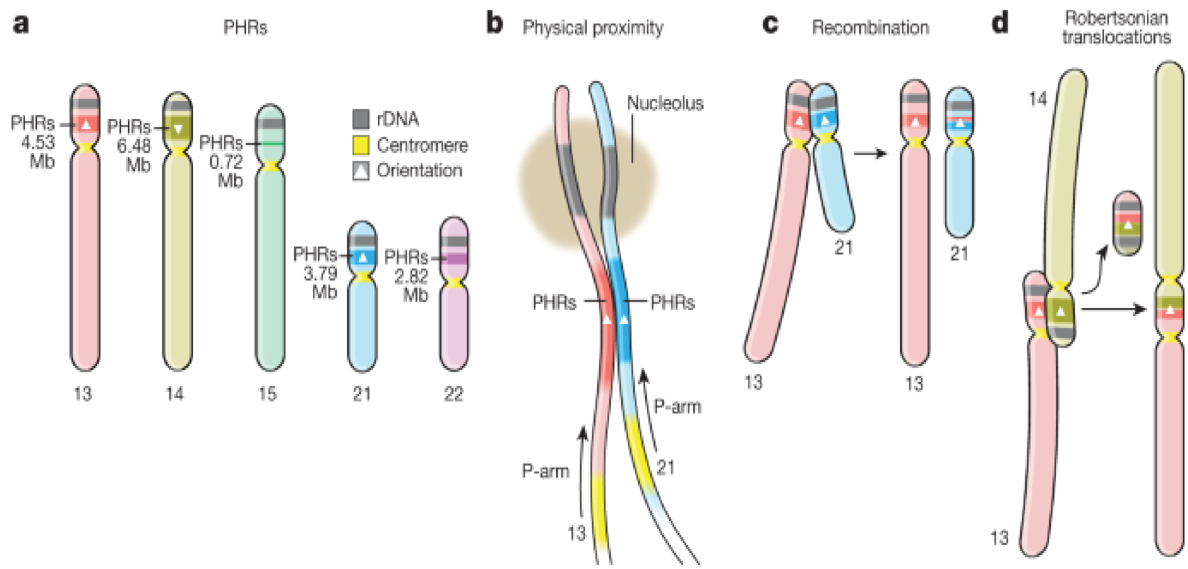
# Discussion

Here we develop multiple lines of evidence indicating active recombination between heterologous human acrocentric chromosomes. First, we find that a symmetric comparison of the sequences of a draft human pangenome contains multi-chromosome communities corresponding to both the sex chromosomes and acrocentric chromosomes (Fig. 1). An acrocentric pangenome graph reveals base-level homologies that outline patterns of

exchange between the heterologous chromosomes (Fig. 2). The graph highlights regions featuring a diverse patchwork of best-match patterns involving non-homologous T2T-CHM13 chromosomes (Fig. 3), and we cross-validate these findings with a T2T diploid assembly of a target sample. We develop an entropy metric sensitive to recombination between heterologous chromosomes—such as that seen between chr. X and chr. Y—and apply it to delineate PHRs where heterologous SAACs recombine (Fig. 4 and Supplementary File 5). Finally, we show that on chr. 13, chr. 14, chr. 21 and chr. 22, the resulting 18 Mb of sequence in the PHRs presents a higher rate of linkage disequilibrium decay than seen in sequences from non-PHR regions of the same chromosomes (Extended Data Fig. 10). These lines of evidence all suggest that heterologous SAACs recombine.

BACs used in a previous cytogenetic study of Robertsonian chromosomes map to the PHRs of chr. 14 and chr. 21, with the recurrent breakpoint region found in a highly homologous region on chr. 13p, chr. 14p and chr. 21p centred on the PRDM9 motif-enriched SST1 array. This leads us to propose that PHRs are maintained by recombination between heterologous chromosomes, which occasionally results in an ROB (Fig. 5). We posit that these homologous regions (Fig. 5a) might share a biological function as sequences proximal to the nucleolar organizing regions. Their proximity (Fig. 5b) can facilitate inter-chromosomal recombination (Fig. 5c)—which may be of both crossover or non-crossover types and may occur during meiosis or mitosis. Owing to an inversion of this region on chr. 14p relative to chr. 13p and chr. 21p, crossover type recombination between pairs chr. 13–chr. 14 and chr. 14–chr. 21 leads to ROBs (Fig. 5d), which our study suggests are a pathological outcome of otherwise benign recombination between heterologous chromosomes.

**Fig. 5: The PHRs of human acrocentric chromosomes.**




**a**, PHRs are found on the rDNA-proximal regions of the SAACs chr. 13, chr. 14, chr. 15, chr. 21 and chr. 22. **b**, PHRs physically co-locate owing to their proximity to the nucleolar organizing regions and rDNA, encouraging sequence exchange. **b**, Patterns of sequence similarity observed in the PHRs indicate ongoing recombination exchange between heterologous chromosomes, in particular chr. 13, chr. 14 and chr. 21, which may be mediated by both non-crossover recombination or crossover of the telomeric ends of heterologous chromosomes. **d**, The PHR surrounding the SST1 arrays on chr. 13, chr. 14 and chr. 21 is nearly identical on all three chromosomes, but is typically inverted on chr. 14 relative to chr. 13 and chr. 21 (triangles). Owing to the inversion, crossover type recombination between PHRs in chr. 14 and chr. 13 or chr. 21 produce an ROB.

The HPRC pangenome provides base-level resolution of homology patterns across many SAAC haplotypes, enabling us to examine in detail the regions in which ROBs occur. We observe that the GC-rich SST1 array lies at the centre of a segmentally duplicated region on chr. 13p, chr. 14p and chr. 21p, which shows a clear pattern of haplotype mixing between these chromosomes. This may also implicate this array as a nucleation point for recombination, as suggested by the observation (in 1 out of 220 oocytes) of heterologous chr. 14p–chr. 21p synapse formation in pachytene oocytes[29]. We speculate that these segmentally duplicated regions are where common ROBs occur, a hypothesis supported by our reanalysis of previous

cytogenetic mapping of the common ROB breakpoint for chr. 14 and chr. 21 (Fig. 4d).

Although we find that SAACs present challenges for assembly methods[4], our validation based on ONT and HiFi data integration shows that the patterns that we observe in HiFi-only assemblies are consistent with ONT reads from the same sample. HG002-Verkko recapitulates key T2T-CHM13-relative untangling patterns also seen in HPRCy1 HiFi assemblies, such as the SST1-linked PHR at chr. 13p11.2 and chr. 21p11.2, rDNA-proximal mixing of all SAACs, and mixing of chr. 22q11.21 and chr. 14q11.2. Our analyses rest on the extensive assembly validation carried out by the HPRC, and observations used to establish signals for recombination are based only on assembly regions deemed to be reliable by mapping analyses[4]. Our study confirms previous hypotheses based on decades of diverse inquiry, which provide additional assurance that patterns observed bioinformatically are biologically grounded. This body of evidence suggests that our definition of the PHRs is likely to evolve with improved resolution of rDNA arrays and distal regions of the SAACs, which remain among the most challenging regions of the human genome to assemble and lie beyond the scope of our presented work.

Our study is fundamentally population-based. We cannot directly observe recombination of SAACs in this context, leaving open questions about recombination mechanisms that may be difficult to resolve from sequence information alone[47]. However, similar to mutation, recombination is a rare event, which makes it easier to measure its distribution over chromosomes in a population genetic context as we have done here. This addresses key issues with many previous studies of recombination in the SAACs, which often feature small numbers of individuals[29,48] selected on the basis of medically relevant genomic states such as trisomy and ROB[49]. Our resolution of the acrocentric PHRs confirms reported homologies between the SAACs[2,3], providing a reference for their structure that will be useful for future genomic and cytogenetic studies. In principle, recombination in the PHRs may be of either crossover or non-crossover type. Our data support both, but outside of recurrent ROBs and our expectation that non-crossover recombination is substantially more common (by a ratio of around 10:1) than crossover recombination[50,51], we lack distinguishing evidence for

either. To estimate the relative rates of each type of event, we can use linkage disequilibrium patterns[50] to study the PHRs in large genomic cohorts[52,53], which will require realigning cohort short read data to T2T-CHM13 or the HPRC pangenome. Future improvements to assembly of the SAACs and the planned increase in the number of individuals included in the HPRC should allow for confident estimates of the relative rates of recombination types.

The co-location of rDNA repeats from different acrocentric chromosomes in a nucleolus provides physical proximity that can facilitate recombination events, both between rDNA repeats and between the adjacent PHRs. Our analyses suggest that the rate of recombination between heterologous pairs of acrocentric chromosomes varies, leading to characteristic patterns in the homology spaces that we have explored. Human cells generally have fewer nucleoli (between one and five) than acrocentric chromosomes[54] (ten). One possibility is that groups of acrocentric chromosomes between which we observe stronger homology and recombination—such as chr. 13, chr. 14 and chr. 21—may be more likely to co-localize to the same nucleolus, as observed in pachytene spermatocytes[15,55]. Proximity, homology, recombination initiation sites and sequence orientation are likely to be factors in the high rate of ROBs between these chromosomes. The HPRC draft human pangenome has enabled us to approach genome evolution from a chromosome scale. By stepping away from a reference-centric model and directly comparing whole-chromosome assemblies of the acrocentric regions, we have obtained sequence-resolved responses to long-standing questions first posed in early cytogenetic studies of human genomes.

# Methods

## Genome assemblies

We analysed the 47 T2T phased diploid de novo assemblies (94 haplotypes in total) produced by the HPRC[4]. We included both T2T-CHM13 version 2[3] and GRCh38.

## Chromosome communities overview

## The homology graph

We first used all-to-all mapping to build a reference-free model of homology relationships in the HPRCy1 pangenome. This models the full HPRCy1 as a mapping graph in which nodes are contigs and edges represent mappings between them. To build the HPRCy1 mapping graph, we generated homology mappings based on chains of 50-kb seeds of 95% average nucleotide identity—which we expect to support homologous recombination[17]—allowing up to $(n - 1) = 93$ alternative mappings over any part of each contig. We first applied this process to map all 38,325 HPRCy1 contigs against all others, obtaining mappings for 38,036 of them covering the 99.9% of the total assembly sequence. This indicated that 38,036 out of 38,325 (99.2%) of the HPRCy1 assembly contigs are homologous to at least 1 other contig. Complex tangles in the assembly graphs used to build the HPRCy1 generate short contigs and tend to result in higher rates of error[4]. Thus, to simplify later analysis and focus on well-resolved regions of the assemblies, we narrowed our focus to consider only mappings involving the 16,118 contigs at least 1 Mb long, covering the 98.72% of the total assembly sequence.

We then built a graph where nodes are contigs and edges represent the mappings between them—the ‘mapping graph’. Edges in this mapping graph have weights equal to the estimated sequence identity multiplied by the length of the mapping. To infer the chromosome represented by each contig, we mapped all contigs against both T2T-CHM13 and GRCh38 references and assigned them a chromosome identity based on this mapping. This mapping graph is very dense, with up to 93 mappings per contig, making it difficult to directly visualize with existing methods. To develop intuition about patterns in this graph, we instead viewed a reduced mapping graph built from the best three mappings per contig segment, labelling each contig with its reference-relative assignment (Fig. 1a). The acrocentric cluster (Fig. 1b) generally matches our prior expectations of higher similarity between chr. 13 and chr. 21, and between chr. 14 and chr. 22[18,19].

## Community detection

To quantify the significance of these patterns, we then applied a community detection algorithm[56] to the full mapping graph. The algorithm assigns each contig to a community such that the total assignment maximizes modularity, which can be understood as the density of (weighted) links inside communities compared to links between communities. This process yielded 31 communities (Supplementary File 1). We hypothesized that each cluster represented one chromosome or chromosome arm. Around half of the chromosomes ($n = 11$) were each represented by a single community. Chromosomes 1, 2, 3, 6 and 18 were each represented in two communities corresponding to their short and long arms, likely due to frequent assembly breaks across their centromeres (Fig. 1c,d). Contigs from chromosomes X and Y fell in the same community, although the short arm of chromosome X was represented in two communities (Fig. 1d). The SAACs formed the community with the most distinct chromosomes and most contigs (1,706 contigs containing 3.91% of HPRCy1 sequence), composed of contigs belonging to the short arms of all the five acrocentric chromosomes plus chr. 21q and chr. 22q (Fig.  1c,d). chr. 13q, chr. 14q and chr. 15q each had their own community. The inclusion of the q-arms of chromosomes 21 and 22 in the community composed of p-arms contigs is likely related to their short lengths compared to chromosomes 13, 14 and 15. We obtained similar results when we increased the sensitivity of the mappings (Supplementary Fig. 29).

In the homology mapping graph of the HPRCy1, only the acrocentric and sex chromosomes form combined communities containing multiple chromosomes. The sex chromosome community reflects the PARs on X and Y[57], which are telomeric regions where these otherwise non-homologous chromosomes recombine as if they were homologues. We hypothesized that the acrocentric community might also reflect ongoing pseudo-homologous recombination

## Community detection workflow

We performed pairwise mapping for all contigs from the 47 T2T phased diploid de novo assemblies with the WFMASH sequence aligner[58] (commit ad8aeba). We set the following parameters:

wfmash HPRCy1.fa -s 50k -l 250k -p 95 -n 93 -Y ’#’ -H 0.001 -m

We used segment seed length of 50 kb (-s), requiring homologous regions at least ~250 kb long (-l) and estimated nucleotide identity of at least ~95% (-p). Having 94 haplotypes in total, we kept up to 93 mappings for each contig (-n). Moreover, we skipped mappings when the query and target had the same prefix before the ‘#’ character (-Y), that is when involving the same haplotype. To properly map through repetitive regions, only the 0.001% of the most frequent kmers were ignored (-H). We skipped the base-level alignment (-m). We also generated pairwise mapping with the same parameters, but using a segment seed length of 10 kb and requiring homologous regions at least ~50 kb long.

From the resulting mappings, we excluded those involving contigs shorter than 1 Mb to reduce the possibility of spurious matches. We then used the paf2net.py Python script (delivered in the PGGB repository) to build a graph representation of the result (a mapping graph), with nodes and edges representing contigs and mappings between them, respectively.

python3 ~/pggb/scripts/paf2net.py -p HPRCy1.1Mbps.paf

The script produces a file representing the edges, a file representing the edge weights, and a file to map graph nodes to sequence names. The weight of an edge is given by the product of the length and the nucleotide identity of the corresponding mapping (higher weights were associated with longer mappings at higher identities). Finally, we used the net2communityes.py Python script (delivered in the PGGB repository) to apply the Leiden algorithm[56], implemented in the igraph tools[59], to detect the underlying communities in the mapping graph.

python3 ~/pggb/scripts/net2communities.py \ -e HPRCy1.1Mbps.edges.list.txt \ -w HPRCy1.1Mbps.edges.weights.txt \ -n HPRCy1.1Mbps.vertices.id2name.chr.txt --accurate-detection

To identify which chromosomes were represented in each community, we partitioned all contigs by mapping them against both T2T-CHM13v1.1 and GRCh38 human reference genomes with WFMASH, this time requiring

homologous regions at least 150 kb long and nucleotide identity of at least 90%.

wfmash chm13+grch38.fa HPRCy1.fa -s 50k -l 150k -p 90 -n 1 -H 0.001 -m -N

We disabled the contig splitting (-N) during mapping to obtain homologous regions covering the whole contigs. For the unmapped contigs, we repeated the mapping with the same parameters, but allowing the contig splitting (without specifying -N). We labelled contigs 'p' or 'q' depending on whether they cover the short arm or the long arm of the chromosome they belonged to. Contigs fully spanning the centromeres were labelled 'pq'. We used such labels to identify the chromosome composition of the communities detected in the mapping graph obtained without reference sequences, and to annotate the nodes in the mapping graph.

To obtain a clean visualization of the homology relationships between the HPRC assemblies, we generated a simpler mapping graph by using the same parameters used for the main graph, but keeping up to 3 mappings for each contig and adding the T2T-CHM13 reference genome version 2, which includes also the complete HG002 chromosome Y (https://www.ncbi.nlm.nih.gov/assembly/GCF_009914755.1):

wfmash HPRCy1+chm13v2.fa -s 50k -l 250k -p 95 -n 3 -Y '#' -H 0.001 -m -w 5000

We set window size for sketching equal to 5000 (-w) to reduce the runtime by sampling fewer kmers. We used the paf2net.py Python script to build the mapping graph and then used Gephi[60] (version 0.9.4) to visualize it. We computed the mapping graph layout by running 'Random Layout' and then the 'Yifan Hu' algorithm.

## Pangenome graph building

For each of the 47 T2T phased diploid de novo assemblies, we mapped all contigs against the T2T-CHM13 human reference genome with the WFMASH sequence aligner (commit ad8aeba). For the HG002 sample, we included two assemblies: the HG002-HPRCy1 phased diploid de novo

assembly (built with HiFi reads) and a phased diploid de novo assembly based on both HiFi and ONT reads, built with the Verkko assembler. We set the following parameters:

```
wfmash chm13.fa assembly.fa -s 50k -l 150k -p 90 -n 1 -H 0.001 -m
```

We used segment seed length of 50 kb (-s), requiring homologous regions at least ~150 kb long (-l) and estimated nucleotide identity of at least ~90% (-p). We kept only one mapping (the best one) for each contig (-n). To properly map through repetitive regions, only the 0.001% of the most frequent kmers were ignored (-H). We skipped the base-level alignment (-m). For the HG002-HPRCy1 contigs, we disabled the contig splitting (-N).

Then, we identified contigs originating from acrocentric chromosomes and covering both the short and long arms of the chromosome they belonged to. We considered only contigs with mappings at least 1 kb long on both arms and at least 1 Mb away from the centromere. We call such contigs 'p–q acrocentric contigs'. For HG002-HPRCy1, only contigs longer than or equal to 300 kb were considered, regardless of covering both arms of the belonging chromosomes.

Finally, we built a pangenome graph with all the p–q acrocentric contigs and both T2T-CHM13 and GRCh38 human reference genomes by applying PGGB[22] (commit a4a6668). We set the following parameters:

```
pggb -i contigs.fa.gz -s 50k -l 250k -p 98 -n 162 -F 0.001 -k 311 -G 13117,13219 -O 0.03
```

We used segment seed length of 50 kb (-s), requiring homologous regions at least ~250 kb long (-l) and estimated nucleotide identity of at least ~98% (-p). Having 142 p–q acrocentric contigs in input (132 from HG002-HPRCy1 and 10 from HG002-Verkko) plus 10 acrocentric chromosomes from the T2T-CHM13 and GRCh38 reference genomes plus 49 HG002-HPRCy1 contigs representing other 10 acrocentric haplotypes (5 maternal and 5 paternal), we kept up to 162 mappings (142 + 10 + 10) for each contig (-n). To properly map through repetitive regions, only the 0.001% of the most frequent kmers were ignored (-F). We filtered out alignment matches shorter than 311 bp to remove possible spurious relationships caused by short

repeated homologies (-k). We set big target sequence lengths and a small sequence padding for two rounds of graph normalization (-G and -O). To visualize the acrocentric pangenome graph, we built the graph layout with ODGI LAYOUT[26] (commit e2de6cb) and visualized with GFAESTUS[61] (commit 50fe37a). This renders sequences and chains of small variants as linear structures, while repeats caused by segmental duplications, inversions and other structural variants tend to form loops.

## Pangenome graph untangling

We untangled the pangenome graph by applying ODGI UNTANGLE (commit e2de6cb). Practically, we projected the graph into an alignment between a set of query (HPRCy1 contigs) and reference (T2T-CHM13) sequences. We set the following parameter:

odgi untangle -i graph.og -e 50000 -m 1000 -n 100 -j 0 -R targets.txt -d cuts.txt

We segmented the graph into regular-sized regions of ~50 kb (-e), merging regions shorter than 1 kb (-m). We reported up to the 100th best target mapping for each query segment (-n), not applying any threshold for the Jaccard similarity (-j). We used all paths in the graphs as queries and projected them against the five acrocentric chromosomes of the T2T-CHM13 genome (-R). Moreover, we emit the cut points used to segment the graph (-d).

For each query segment, if there were multiple best hits against different targets (that is, hits with the same, highest Jaccard similarity), we put as the first one the hit having as target the chromosome of origin of the query (obtained from the chromosome partitioning of the contigs).

We repeated the graph untangling another five times, but constrained the algorithm to use only one of the acrocentric chromosomes of T2T-CHM13 as a target at a time (-r) and return the best-matching hit (-n).

odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr13 -c cuts.txt odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr14 -c cuts.txt odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr15 -c cuts.txt odgi

untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr21 -c cuts.txt odgi untangle -i graph.og -e 50000 -m 1000 -n 1 -j 0 -r chr22 -c cuts.txt

We used the cut points generated when using all of the acrocentric chromosomes of T2T-CHM13 as targets (-c). In this way, all untangling runs (six in total) used the same cut points for the segment boundaries.

Finally, we 'grounded' the untangled output generated with all acrocentric chromosomes as targets: in more detail, each untangled query segment was placed against a particular acrocentric chromosome (not only the best-matching one) by using the untangled outputs constrained to a single target. We split the result by acrocentric chromosome and kept only queries untangling both p- and q-arms of the targets. Furthermore, we removed query segments overlapping regions flagged in the assemblies as unreliable (that is, having coverage issues) by FLAGGER[4]. FLAGGER is a HiFi read-based pipeline that detects different types of mis-assemblies within a phased diploid assembly by identifying read-mapping coverage inconsistencies across the maternal and paternal haplotypes. To focus on the more similar query-target hits, we used the Jaccard metric to estimate the sequence identity by applying the corrected formula reported in ref. [62], and retained only results at greater than 90% estimated identity. To analyse the orientation status of HPRCy1 contigs in the segmental duplication centred on the SST1 array, we generated a new pangenome graph with ODGI FLIP (commit 0b21b35). In more detail, we first flipped paths around if they tend to be in the reverse complement orientation relative to the pangenome graph. This leads to having a uniform orientation for the HPRCy1 contigs, all in forward orientation with respect to the graph. Then, we untangled the flipped graph in the same way as described above. We displayed the untangling results for each acrocentric chromosome with the R development environment (version 3.6.3), equipped with the following packages: tidyverse (version 1.3.0), RColorBrewer (version 1.1.2), ggplot2 (version 3.3.3) and ggrepel (version 0.9.1).

## Recombination pattern analysis

### Aggregating best-hit untangle results

For each group of HPRCy1-acro contigs anchored to a T2T-CHM13 acrocentric q-arm, we counted the number of contigs having as best-hit each one of the acrocentrics. In particular, for each base position of the T2T-CHM13 acrocentric chromosome of each group, we quantified how many times each of the acrocentrics appeared as best-hit in the pangenome graph untangling. We considered only best hits with an estimated identity of at least 90%.

**Regional homology entropy**

To quantify the degree of disorder in the untangling result, we calculated the diversity entropy across the different acrocentrics that were present as best-hit. In more detail, we projected each HPRCy1 acrocentric p–q contig against the T2T-CHM13 acrocentric to which it is anchored via the q-arm and associated each reference base position to the corresponding acrocentric best-hit. We considered only best hits with an estimated identity of at least 90%. Then, we computed the Shannon diversity index (SDI) in windows 50 kb long over the contigs. We used −1 as missing SDI value in the regions where the contigs do not match any targets. For each group of contigs, we aggregated the SDI results by computing their average (ignoring the missing SDI values) for each reference base position. We call this metric the positional homology entropy, and it serves to show regions where contigs can be described as mosaics of different reference chromosomes. However, it cannot distinguish regions where there are different orders of reference chromosome similarity—that might be indicative of recombination exchange—from those where there is regional diversity in each contig's relationship to T2T-CHM13. The latter case could occur if T2T-CHM13 itself contains rare recombinations between acrocentrics, or where ancient homology might result in 'noise' in untangling alignments as contigs pick from two equally good alternative mappings. To avoid these pitfalls and establish a more stringent graph-space recombination metric, we then extended the untangling diversity metric to operate on multiple mappings.

**Positional homology entropy**

To take into account the other hits in addition to the first one, including their order, we generalized the diversity entropy metric to work over orders of the

top 5 untangling hits and consider all contigs jointly. For each reference segment, we collected the corresponding best 5 untangling hits for each of the HPRCy1 acrocentric p–q contig; this is possible because the reference segments are stable across all contigs. We considered only best hits with an estimated identity of at least 90%. To avoid driving the untangle entropy by intra-chromosomal similarity caused by segmental duplications modelled in the structure of the PVG (as seen in loops on chr. 13q, chr. 15q and chr. 22q; Fig. 2), we ignored consecutive duplicate target hits—in other words, we took the ordered set of unique reference targets. When multiple contig segments were grounded against the same reference segment, we considered the first contig segment having the best grounding, that is having the highest estimated identity when placed against the current reference segment. Then, we ranked the five best hits by estimated similarity. Finally, for each reference segment, we computed the SDI across all available five best hits orders. We used −1 as missing SDI value in the reference regions without any contig matches. We kept in the output also the information about how many HPRCy1 acrocentric p–q contigs contributed to the entropy computation in each reference segment. This yielded the positional homology entropy.

## PHR derivation

To obtain the PHRs, we aggregated the final results by considering regions with positional homology entropy greater than 0 and supported by at least 1 contig, merging with BEDtools[63] those that were less than 30 kb away, and removing merged regions shorter than 30 kb.

## Display of untangle mosaics

We displayed the aggregated results for each acrocentric chromosome. We used genome annotations for the first 25 Mb of each acrocentric chromosome, using T2T-CHM13v2.0 UCSC trackhub (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hub_3671779_hs1). We made the figures with the scripts available at https://github.com/pangenome/chromosome_communities/tree/main/scripts. To plot the figures, we used R (version 3.6.3), equipped with the following packages: tidyverse (version 1.3.0), RColorBrewer (version 1.1.2), ggplot2

(version 3.3.3) and ggrepel (version 0.9.1). Finally, we used Inkscape (https://inkscape.org/) to compose main text figures based on the results, and provide supplementary figures directly produced by these methods.

## HPRCy1 SNP density plots

We displayed biallelic SNP density in the full HPRCy1 draft pangenome built with PGGB[4] versus both GRCh38 and T2T-CHM13. To do so, we extracted biallelic SNPs from the released VCF files versus each chromosome, for both references (get_bisnp.sh). Because T2T-CHM13 version 1.1, which was used in HPRCy1, does not have a Y chromosome, we used that of GRCh38, which includes masked PAR1 and PAR2 regions. We displayed biallelic SNP density in bins of 100 kb, using R (version 4.1.1) and tidyverse (version 1.3.1) package (plot_bisnp_dens.R).

## ROB breakpoints

We mapped BAC clones from ref. [8] against the T2T-CHM13 human reference genome with the WFMASH sequence aligner (commit ad8aeba). We kept only mappings covering acrocentric chromosomes and with an estimated identity of at least 90%. To plot the figures, we used R (version 3.6.3), equipped with the following packages: tidyverse (version 1.3.0), RColorBrewer (version 1.1.2), ggplot2 (version 3.3.3) and ggrepel (version 0.9.1). We coloured BAC clones' mappings according to ref. [8].

## Maximum likelihood phylogenetic analysis

We conducted the phylogenetic analysis by using the maximum likelihood method based on the best-fit substitution model (Kimura 2-parameter +G, parameter = 5.5047) inferred by Jmodeltest2[64] with 1,000 bootstrap replicates. Bootstrap values higher than 75 are indicated at the base of each node.

## Recombination hotspots analysis

We obtained the human PRDM9 binding motifs (17 in total) from ref. [41] and used FIMO[65] to scan their occurrences in T2T-CHM13v2.0 human reference genome:

```
fimo --thresh 1.0E-4 PRDM9_motifs.human.txt chm13v2.fa
```

FIMO computes a log-likelihood ratio score for each motif with respect to each sequence position and converts these scores to *P* values using dynamic programming (assuming a zero-order null model in which sequences are generated at random with user-specified per-letter background frequencies) and then estimate false discovery rates[65]. Each motif is associated with a measure of how likely it represents a true binding target for PRDM9. We retained for downstream analyses only motifs for which such a measure is at least 70% (14 of 17). For each motif, we counted the number of occurrences present in windows 20 kb long across each T2T-CHM13v2.0 chromosome by using BEDtools[63].

```
bedtools intersect -a chm13v2.windows_20kbp.bed -b
fimo_output.motif$i.bed
```

To plot the figures, we used R (version 3.6.3), equipped with the following packages: tidyverse (version 1.3.0), RColorBrewer (version 1.1.2), ggplot2 (version 3.3.3) and ggrepel (version 0.9.1).

## Linkage disequilibrium analysis

We identified variants embedded in the pangenome graph by using VG DECONSTRUCT[5]:

```
vg deconstruct -P chm13 -H ‘?’ --ploidy 1 -e -a graph.gfa > variants.vcf
```

We called variants with respect to the T2T-CHM13 reference genome (-P), reporting variants for each HPRCy1 acrocentric p–q contig (-H and --ploidy). We considered only traversals that correspond to paths (that is, contigs) in the graph (-e) and also reported nested variation (-a). From the variant set, we considered only single nucleotide variants. We estimated linkage disequilibrium between pairs of markers within 70 kb by using PLINK v1.9[66] upon specification of haploid sets and retaining all values of

$r^2 > 0$ (plot_ld_1.R). Finally, we generated binned linkage disequilibrium decay plots with confidence intervals using R (version 3.6.3), focusing on pairs less than 4 kb apart.

## Validating homology mosaicism

### HG002-Verkko assembly

We applied an earlier version of Verkko (beta 1, commit vd3f0b941b5facf5807c303b0c0171202d83b7c74) to build a diploid assembly graph for the HG002 cell line using the HiFi (105x) and ONT (85x) reads as described[67]. The resulting assembly graph resolves the proximal junction in single contigs for each haplotype up to multi-mega bases, while the distal junctions remain to be resolved. We used homopolymer compressed markers from the parental Illumina reads to assign unitigs to maternal, paternal haplotype or ambiguous when not enough markers supported either haplotype. For estimating the number of times a unitig has to be visited, we aligned HiFi and ONT reads to the assembly graph using GraphAligner with the following parameters: –seeds-mxm-length 30–seeds-mem-count 10000 -b 15–multimap-score-fraction 0.99–precise-clipping 0.85–min-alignment-score 5000–clip-ambiguous-ends 100–overlap-incompatible-cutoff 0.15–max-trace-count 5–hpc-collapse-reads–discard-cigar[68]. Four distal junctions were connected to the rDNA arrays with ambiguous nodes connecting the maternal and paternal nodes, supporting they belong to the same chromosome. Two distal junction unitigs, one maternal and paternal, were disconnected from each other but connected to the rDNA arrays, which were assigned to the same chromosome. Using the marker and ONT alignments, we identified paths in the graph and assigned them according to the most supported haplotype. If only ambiguous nodes were present between the haplotype assigned unitigs, with no ONT reads to resolve the path, nodes were randomly assigned to one haplotype to build the contig. After all paths were identified, we produced the consensus using verkko --assembly <path-to-original-assembly> --paths <path-to-paths>. The entire procedure to produce parental markers, tagging unitigs according to its haplotype on the assembly graph and finding the path using ONT reads is now available in the latest Verkko (v1) in a more automated way.

## Untangling validation

To provide cross-validation of the HiFi contig assemblies and our analysis of them, we compared the untangling of two assemblies of the same sample (HG002). One was made with the HPRCy1 pipeline, while the other used the Verkko diploid T2T assembler. Verkko employs ONT to untangle ambiguous regions in a HiFi-based assembly graph, automating techniques first developed in production of T2T-CHM13. Verkko's assembly aggregates information from ONT, thus providing a single integrated target for cross-validation of our analysis using an alternative sequencing and assembly approach.

We validated the results of the pangenome untangling by comparing the best hits of the two HG002 assemblies, the one built with HiFi reads and the other based on both HiFi and ONT reads, built with the Verkko assembler[67]. For each base position of each T2T-CHM13 acrocentric chromosome, we compared the untangling best-hit of the HG002-HPRCy1 contigs with the best-hit supported by HG002-Verkko contigs. We considered only best hits with an estimated identity of at least 90%. We defined reference regions as concordant when both HG002 assemblies supported the same T2T-CHM13 acrocentric as best-hit. We treated the two haplotypes (maternal and parental) separately.

We observe a high degree of concordance between the two methods at a level of the chromosome homology mosaicism plots. The best-hit untangling shows similar patterns in the HG002-Verkko assembly as those seen in HG002-HPRCy1 (Supplementary Figs. 9–13). However, some SAAC haplotypes appear to be poorly assembled in HG002-HPRCy1. On the q-arms we measured 99.93% concordance between HG002-HPRCy1 and HG002-Verkko untangling results, but only 87.45% concordance on the p-arms (Supplementary File 8). This lower level is consistent with greater difficulty in assembling the SAACs due to their multiple duplicated sequences (including the PHRs), satellite arrays and the rDNA. We found the discordance was driven by a single chromosome haplotype: while most p-arms achieve around 90% concordance, HG002-HPRCy1 14p-maternal exhibits a high degree of discordance in the assemblies (66.19% concordance) (Supplementary Fig. 19). Although this $n = 1$ validation

focuses on only 10 haplotypes, it incorporates many independent reads provided by deep HiFi (105x) and ONT (85x) data used in HG002-Verkko. We thus have compared the concordance between structures observed in single molecule reads across the SAACs and the HG002-HPRCy1 assembly that represents the HiFi-only assembly process that produced our pangenome.

However, this analysis should be seen as presenting a lower bound on our process accuracy. We are considering *all* HG002-HPRCy1 contigs that map to the acrocentrics—not only those that would meet our centromere-crossing requirement, and HG002-HPRCy1 is itself more fragmented than the other assemblies which we have selected for the acro-PVG[4]. The fragmented nature of its contigs (only one from chr. 22 meets our p–q mapping requirement) may introduce additional disagreements with HG002-Verkko. The overall result indicates that most patterns observed in HiFi-only assemblies are likely to be supported by an automated near-T2T assembly of the same sample.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Assemblies produced by the HPRC are available at AnVIL (https://anvilproject.org/), in the AnVIL_HPRC workspace. Data are also available as part of the AWS Open Data Program (https://registry.opendata.aws/) in the human-pangenomics S3 bucket (https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html). In addition, the data have been uploaded to the International Nucleotide Sequence Database Collaboration (INSDC) for long-term storage and availability. Supporting information about the data (including index files with S3 and GCP file locations) can be found at the following GitHub repository: https://github.com/human-pangenomics/HPP_Year1_Assemblies. All supplementary files, including the PVG and its layout, are available on Zenodo at https://doi.org/10.5281/zenodo.7692554.

# Code availability

Code and links to methods and tools used to perform all the analyses and produce all the figures can be found on Zenodo at https://doi.org/10.5281/zenodo.7697614.

# References


1. Floutsakou, I. et al. The shared genomic architecture of human nucleolar organizer regions. *Genome Res.* **23**, 2003–2012 (2013).

2. van Sluis, M. et al. Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions. *Genes Dev.* **33**, 1688–1701 (2019).

3. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

4. Liao, W.-W. et al. A draft human pangenome reference. *Nature* https://doi.org/10.1038/s41586-023-05896-x (2023).

5. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).

6. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

7. Huttley, G. A., Smith, M. W., Carrington, M. & O'Brien, S. J. A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722 (1999).

8. Jarmuz-Szymczak, M., Janiszewska, J., Szyfter, K. & Shaffer, L. G. Narrowing the localization of the region breakpoint in most frequent Robertsonian translocations. *Chromosome Res.* **22**, 517–532 (2014).

9. Hamerton, J. L., Canning, N., Ray, M. & Smith, S. A cytogenetic survey of 14,069 newborn infants. I. Incidence of chromosome abnormalities. *Clin. Genet.* **8**, 223–243 (1975).

10. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

11. Mack, H. & Swisshelm, K. in *Brenner's Encyclopedia of Genetics* 2nd edn (eds Maloy, S. & Hughes, K.) 301–305 (Academic Press, 2013).

12. Spinner, N. B. in *Brenner's Encyclopedia of Genetics* 2nd edn (eds Maloy, S. & Hughes, K.) 546–548 (Academic Press, 2013).

13. Lindström, M. S. et al. Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene* **37**, 2351–2366 (2018).

14. Kobayashi, T. Regulation of ribosomal RNA gene copy number and its role in modulating genome integrity and evolutionary adaptability in yeast. *Cell. Mol. Life Sci.* **68**, 1395–1403 (2011).

15. Holm, P. B. & Rasmussen, S. W. Human meiosis I. The human pachytene karyotype analyzed by three dimensional reconstruction of the synaptonemal complex. *Carlsberg Res. Commun.* **42**, 283 (1977).

16. Choo, K. H., Vissel, B., Brown, R., Filby, R. G. & Earle, E. Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations. *Nucleic Acids Res.* **16**, 1273–1284 (1988).

17. Peng, Z. et al. Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome. *Hum. Mol. Genet.* **24**, 1225–1233 (2015).

18. Greig, G. M., Warburton, P. E. & Willard, H. F. Organization and evolution of an alpha satellite DNA subset shared by human

chromosomes 13 and 21. *J. Mol. Evol.* **37**, 464–475 (1993).

19. Jørgensen, A. L., Kølvraa, S., Jones, C. & Bak, A. L. A subfamily of alphoid repetitive DNA shared by the NOR-bearing human chromosomes 14 and 22. *Genomics* **3**, 100–109 (1988).

20. Cotter, D. J., Brotman, S. M. & Wilson Sayres, M. A. Genetic diversity on the human X chromosome does not support a strict pseudoautosomal boundary. *Genetics* **203**, 485–492 (2016).

21. Ross, M. T. et al. The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).

22. Garrison, E. & Guarracino, A. et al. Building pangenome graphs. Preprint at *bioRxiv* 2023.04.05.535718 https://doi.org/10.1101/2023.04.05.535718 (2023).

23. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).

24. Eizenga, J. M. et al. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).

25. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).

26. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, 3319–3326 (2022).

27. Marco-Sola, S. et al. Optimal gap-affine alignment in $O(s)$ space. *Bioinformatics* **39**, btad074 (2023).

28. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *Bioinformatics* **39**, btac743 (2023).

29. Cheng, E. Y. & Naluai-Cecchini, T. FISHing for acrocentric associations between chromosomes 14 and 21 in human oogenesis. *Am.*

*J. Obstet. Gynecol.* **190**, 1781–5 (2004).

30. Epstein, N. D. et al. A new moderately repetitive DNA sequence family of novel organization. *Nucleic Acids Res.* **15**, 2327–2341 (1987).

31. Hoyt, S. J. et al. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).

32. Tremblay, D. C., Alexander, G. Jr, Moseley, S. & Chadwick, B. P. Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC Genomics* **11**, 632 (2010).

33. González, B. et al. Somatic hypomethylation of pericentromeric SST1 repeats and tetraploidization in human colorectal cancer cells. *Cancers* **13**, 5353 (2021).

34. Henderson, A. S., Warburton, D. & Atwood, K. C. Ribosomal DNA connectives between human acrocentric chromosomes. *Nature* **245**, 95–97 (1973).

35. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).

36. Kinene, T., Wainaina, J., Maina, S. & Boykin, L. M. in *Encyclopedia of Evolutionary Biology* (ed. Kliman, R. M.) 489–493 (Academic Press, 2016).

37. Arenas, M. The importance and application of the ancestral recombination graph. *Front. Genet.* **4**, 206 (2013).

38. Veerappa, A. M., Padakannaya, P. & Ramachandra, N. B. Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. *Funct. Integr. Genomics* **13**, 285–293 (2013).

39. Paigen, K. & Petkov, P. M. PRDM9 and its role in genetic recombination. *Trends Genet.* **34**, 291–300 (2018).

40. Zickler, D. & Kleckner, N. Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harb. Perspect. Biol.* **7**, a016626 (2015).

41. Altemose, N. et al. A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *eLife* **6**, e28383 (2017).

42. Nambiar, M. & Smith, G. R. Repression of harmful meiotic recombination in centromeric regions. *Semin. Cell Dev. Biol.* **54**, 188–197 (2016).

43. Paten, B. et al. Superbubbles, ultrabubbles, and cacti. *J. Comput. Biol.* **25**, 649–663 (2018).

44. Beichman, A. C., Phung, T. N. & Lohmueller, K. E. Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3* **7**, 3605–3620 (2017).

45. Bosch, E. et al. Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: most population isolates do not show increased LD. *BMC Genomics* **10**, 338 (2009).

46. Roberts, P. A. Difference in the behaviour of eu- and hetero-chromatin: crossing-over. *Nature* **205**, 725–726 (1965).

47. Ahuja, J. S., Harvey, C. S., Wheeler, D. L. & Lichten, M. Repeated strand invasion and extensive branch migration are hallmarks of meiotic recombination. *Mol. Cell* **81**, 4258–4270.e4 (2021).

48. Guissani, U., Facchinetti, B., Cassina, G. & Zuffardi, O. Mitotic recombination among acrocentric chromosomes' short arms. *Ann. Hum. Genet.* **60**, 91–97 (1996).

49. Bandyopadhyay, R. et al. Mosaicism in a patient with Down syndrome reveals post-fertilization formation of a Robertsonian translocation and isochromosome. *Am. J. Med. Genet. A* **116A**, 159–163 (2003).

50. Gay, J., Myers, S. & McVean, G. Estimating meiotic gene conversion rates from population genetic data. *Genetics* **177**, 881–894 (2007).

51. Cole, F., Keeney, S. & Jasin, M. Comprehensive, fine-scale dissection of homologous recombination outcomes at a hot spot in mouse meiosis. *Mol. Cell* **39**, 700–710 (2010).

52. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

53. Chen, S. et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. Preprint at *bioRxiv* 2022.03.20.485034 https://doi.org/10.1101/2022.03.20.485034 (2022).

54. Berríos, S. et al. Number and nuclear localisation of nucleoli in mammalian spermatocytes. *Genetica* **121**, 219–228 (2004).

55. Berríos, S. & Fernández-Donoso, R. Nuclear architecture of human pachytene spermatocytes: quantitative analysis of associations between nucleolar and XY bivalents. *Hum. Genet.* **86**, 103–116 (1990).

56. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

57. Helena Mangs, A. & Morris, B. J. The human pseudoautosomal region (PAR): origin, function and future. *Curr. Genomics* **8**, 129–136 (2007).

58. Guarracino, A., Mwaniki, N., Marco-Sola, S. & Garrison, E. wfmash: a pangenome-scale pairwise aligner. *Zenodo* https://doi.org/10.5281/zenodo.6949373 (2021).

59. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Int. J. Complex Syst.* 1695 (2006).

60. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. *ICWSM* **3**, 361–362 (2009).

61. Fischer, C. & Garrison, E. chfi/gfaestus: a pangenome graph browser. *Zenodo* https://doi.org/10.5281/zenodo.6954036 (2022).

62. Belbasi, M., Blanca, A., Harris, R. S., Koslicki, D. & Medvedev, P. The minimizer Jaccard estimator is biased and inconsistent. *Bioinformatics* **38**, i169–i176 (2022).

63. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

64. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).

65. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

66. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

67. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01662-6 (2023).

68. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).

# Acknowledgements

Our work depends on the HPRC draft human pangenome resource established in the accompanying Article[4], and we thank the production and assembly groups for their efforts in establishing this resource. This work used the computational resources of the UTHSC Octopus cluster and NIH HPC Biowulf cluster. We acknowledge support in maintaining these systems that was critical to our analyses. The authors thank M. Miller for the development of a graphical synopsis of our study (Fig. 5); and R. Williams and N. Soranzo for support and guidance in the design and discussion of our

work. This work was supported, in part, by National Institutes of Health/NIDA U01DA047638 (E.G.), National Institutes of Health/NIGMS R01GM123489 (E.G.), NSF PPoSS Award no. 2118709 (E.G. and C.F.), the Tennessee Governor's Chairs programme (C.F. and E.G.), National Institutes of Health/NCI R01CA266339 (T.P., L.G.d.L. and J.L.G.), and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (A.R., S.K. and A.M.P.). We acknowledge support from Human Technopole (A.G.), Consiglio Nazionale delle Ricerche, Italy (S.B. and V.C.), and Stowers Institute for Medical Research (T.P., L.G.d.L., B.R. and J.L.G.).

## Author information

### Authors and Affiliations

1. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

   Andrea Guarracino, Christian Fischer, Pjotr Prins, Flavia Villani, Vincenza Colonna & Erik Garrison

2. Genomics Research Centre, Human Technopole, Milan, Italy

   Andrea Guarracino

3. Institute of Genetics and Biophysics, National Research Council, Naples, Italy

   Silvia Buonaiuto & Vincenza Colonna

4. Stowers Institute for Medical Research, Kansas City, MO, USA

   Leonardo Gomes de Lima, Tamara Potapova, Boris Rubinstein & Jennifer L. Gerton

5. Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Arang Rhie, Sergey Koren, Ann McCartney, Sergey Nurk, Mikko Rautiainen, Brian Walenz & Adam M. Phillippy

6. Division of Oncology, Department of Internal Medicine, Washington University School of Medicine, St Louis, MO, USA

   Haley J. Abel

7. McDonnell Genome Institute, Washington University School of Medicine, St Louis, MO, USA

   Lucinda L. Antonacci-Fulton, Sarah Cody, Robert S. Fulton, Wen-Wei Liao, Allison A. Regier & Chad Tomlinson

8. UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA

   Mobin Asri, Xian H. Chang, Mark Diekhans, Jordan M. Eizenga, Marina Haukness, David Haussler, Glenn Hickey, Julian K. Lucas, Charles Markello, Karen H. Miga, Jean Monlong, Adam M. Novak, Hugh E. Olsen, Benedict Paten, Trevor Pesout & Jouni Sirén

9. Google, Mountain View, CA, USA

   Gunjan Baid, Anastasiya Belyaeva, Andrew Carroll, Pi-Chuan Chang, Daniel E. Cook, Alexey Kolesnikov, Maria Nattestad & Kishwar Shafin

10. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA

    Carl A. Baker, Evan E. Eichler, William T. Harvey, Kendra Hoekzema, Jennifer Kordosky, Alexandra P. Lewis, Katherine M. Munson, David Porubsky & Mitchell R. Vollger

11. European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK

Konstantinos Billis, Susan Fairley, Paul Flicek, Adam Frankish, Carlos Garcia Giron, Leanne Haggerty, Thibaut Hourlier, Fergal J. Martin & Francesca Floriana Tricomi

12. Department of Human Genetics, McGill University, Montreal, Quebec, Canada

Guillaume Bourque

13. Canadian Center for Computational Genomics, McGill University, Montreal, Quebec, Canada

Guillaume Bourque

14. Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan

Guillaume Bourque

15. Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA

Mark J. P. Chaisson & Tsung-Yu Lu

16. Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

Haoyu Cheng, Justin Chu, Xiaowen Feng & Heng Li

17. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Haoyu Cheng, Xiaowen Feng & Heng Li

18. Arizona State University, Barrett and O'Connor Washington Center, Washington, DC, USA

Robert M. Cook-Deegan

19. School of Biological Sciences, Washington State University, Pullman, WA, USA

Omar E. Cornejo

20. Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Daniel Doerr, Peter Ebert, Jana Ebler, Hugo Magalhães, Pierre Marijon & Tobias Marschall

21. Howard Hughes Medical Institute, Chevy Chase, MD, USA

Evan E. Eichler, David Haussler & Erich D. Jarvis

22. The Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA

Olivier Fedrigo, Giulio Formenti & Jacquelyn Mountcastle

23. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Adam L. Felsenfeld, Baergen I. Schultz, Michael W. Smith & Heidi J. Sofia

24. Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

Yan Gao

25. NNF Center for Biosustainability, Technical University of Denmark, Copenhagen, Denmark

Shilpa Garg

26. Institute for Society and Genetics, College of Letters and Science, University of California, Los Angeles, Los Angeles, CA, USA

Nanibaa’ A. Garrison

27. Institute for Precision Health, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

Nanibaa’ A. Garrison

28. Division of General Internal Medicine and Health Services Research, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

Nanibaa’ A. Garrison

29. Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, USA

Richard E. Green

30. Dovetail Genomics, Scotts Valley, CA, USA

Richard E. Green

31. Quantitative Life Sciences, McGill University, Montreal, Québec, Canada

Cristian Groza

32. Department of Genetics, Yale University School of Medicine, New Haven, CT, USA

Ira Hall, Wen-Wei Liao & Shuangjia Lu

33. Center for Genomic Health, Yale University School of Medicine, New Haven, CT, USA

Ira Hall

34. Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany

Simon Heumos

35. Biomedical Data Science, Department of Computer Science, University of Tübingen, Tübingen, Germany

Simon Heumos

36. Tree of Life, Wellcome Sanger Institute, Cambridge, UK

Kerstin Howe & Jonathan M. D. Wood

37. Northeastern University, Boston, MA, USA

Miten Jain

38. The Rockefeller University, New York, NY, USA

Erich D. Jarvis

39. Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

Hanlee P. Ji & HoJoon Lee

40. Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Eimear E. Kenny

41. Program in Bioethics and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA

Barbara A. Koenig

42. Genome Biology Unit, European Molecular Biology LaboratoryGenome Biology Unit, Heidelberg, Germany

Jan O. Korbel

43. Department of Medicine, Washington University School of Medicine, St Louis, MO, USA

Wen-Wei Liao

44. Computer Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain

Santiago Marco-Sola

45. Departament d'Arquitectura de Computadors i Sistemes Operatius, Universitat Autònoma de Barcelona, Barcelona, Spain

Santiago Marco-Sola

46. Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Tobias Marschall

47. Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

Jennifer McDaniel, Nathan D. Olson, Justin Wagner & Justin M. Zook

48. Coriell Institute for Medical Research, Camden, NJ, USA

Matthew W. Mitchell

49. Department of Computer Science, University of Pisa, Pisa, Italy

Moses Njagi Mwaniki

50. Department of Public Health Sciences, University of California, Davis, Davis, CA, USA

Alice B. Popejoy

51. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

Daniela Puiu & Aleksey V. Zimin

52. Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA, USA

Samuel Sacco

53. Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany

Ashley D. Sanders

54. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Valerie A. Schneider & Françoise Thibaud-Nissen

55. Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark

Jonas A. Sibbesen

56. Al Jalila Genomics Center of Excellence, Al Jalila Children's Specialty Hospital, Dubai, United Arab Emirates

Ahmad N. Abou Tayoun

57. Center for Genomic Discovery, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates

Ahmad N. Abou Tayoun

58. Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA, USA

Mitchell R. Vollger

59. Department of Genetics, Washington University School of Medicine, St Louis, MO, USA

Ting Wang

60. Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA

Aleksey V. Zimin

## Consortia

## Human Pangenome Reference Consortium


- Haley J. Abel
- , Lucinda L. Antonacci-Fulton
- , Mobin Asri
- , Gunjan Baid
- , Carl A. Baker
- , Anastasiya Belyaeva
- , Konstantinos Billis
- , Guillaume Bourque
- , Silvia Buonaiuto
- , Andrew Carroll
- , Mark J. P. Chaisson
- , Pi-Chuan Chang
- , Xian H. Chang
- , Haoyu Cheng
- , Justin Chu
- , Sarah Cody
- , Vincenza Colonna
- , Daniel E. Cook
- , Robert M. Cook-Deegan
- , Omar E. Cornejo
- , Mark Diekhans
- , Daniel Doerr
- , Peter Ebert
- , Jana Ebler
- , Evan E. Eichler
- , Jordan M. Eizenga

- , Susan Fairley
- , Olivier Fedrigo
- , Adam L. Felsenfeld
- , Xiaowen Feng
- , Christian Fischer
- , Paul Flicek
- , Giulio Formenti
- , Adam Frankish
- , Robert S. Fulton
- , Yan Gao
- , Shilpa Garg
- , Erik Garrison
- , Nanibaa' A. Garrison
- , Carlos Garcia Giron
- , Richard E. Green
- , Cristian Groza
- , Andrea Guarracino
- , Leanne Haggerty
- , Ira Hall
- , William T. Harvey
- , Marina Haukness
- , David Haussler
- , Simon Heumos
- , Glenn Hickey
- , Kendra Hoekzema
- , Thibaut Hourlier
- , Kerstin Howe
- , Miten Jain
- , Erich D. Jarvis
- , Hanlee P. Ji
- , Eimear E. Kenny
- , Barbara A. Koenig
- , Alexey Kolesnikov
- , Jan O. Korbel
- , Jennifer Kordosky
- , Sergey Koren
- , HoJoon Lee

- , Alexandra P. Lewis
- , Heng Li
- , Wen-Wei Liao
- , Shuangjia Lu
- , Tsung-Yu Lu
- , Julian K. Lucas
- , Hugo Magalhães
- , Santiago Marco-Sola
- , Pierre Marijon
- , Charles Markello
- , Tobias Marschall
- , Fergal J. Martin
- , Ann McCartney
- , Jennifer McDaniel
- , Karen H. Miga
- , Matthew W. Mitchell
- , Jean Monlong
- , Jacquelyn Mountcastle
- , Katherine M. Munson
- , Moses Njagi Mwaniki
- , Maria Nattestad
- , Adam M. Novak
- , Sergey Nurk
- , Hugh E. Olsen
- , Nathan D. Olson
- , Benedict Paten
- , Trevor Pesout
- , Adam M. Phillippy
- , Alice B. Popejoy
- , David Porubsky
- , Pjotr Prins
- , Daniela Puiu
- , Mikko Rautiainen
- , Allison A. Regier
- , Arang Rhie
- , Samuel Sacco
- , Ashley D. Sanders

- , Valerie A. Schneider
- , Baergen I. Schultz
- , Kishwar Shafin
- , Jonas A. Sibbesen
- , Jouni Sirén
- , Michael W. Smith
- , Heidi J. Sofia
- , Ahmad N. Abou Tayoun
- , Françoise Thibaud-Nissen
- , Chad Tomlinson
- , Francesca Floriana Tricomi
- , Flavia Villani
- , Mitchell R. Vollger
- , Justin Wagner
- , Brian Walenz
- , Ting Wang
- , Jonathan M. D. Wood
- , Aleksey V. Zimin
- & Justin M. Zook

## Contributions

Paper writing: A.G. and E.G. Paper editing: A.G., S.B., L.G.d.L., T.P., A.R., S.K., B.R., C.F., J.L.G., A.M.P., V.C. and E.G. Development of algorithms and software: A.G. and E.G. Chromosome community detection: A.G. and E.G. Pangenome graph building and analyses: A.G. and E.G. Pangenome visualization: A.G., C.F. and E.G. ROB breakpoints: A.G., T.P., J.L.G. and E.G. Maximum likelihood phylogenetic analysis: L.G.d.L. Recombination hotspots analysis: A.G. Linkage disequilibrium analysis: A.G., S.B. and V.C. Homology mosaicism analysis: A.G. and E.G. Physical distance modelling: B.R. and J.L.G. HG002-Verkko assembly: A.R., S.K. and A.M.P. Untangling validation: A.G. and E.G.

## Corresponding author

Correspondence to Erik Garrison.

# Ethics declarations

## Competing interests

The authors declare no competing interests.

# Peer review

## Peer review information

*Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

# Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# Extended data figures and tables

## Extended Data Fig. 1 (A) Evolutionary strata 5 and 4.

Visualization with Saffire (https://mrvollger.github.io/SafFire/) of the alignment between T2T-CHM13 X and Y reveals that strata 5 and 4 feature low identity (~90%), numerous inversions, and some rearrangements; (**B**) X chromosome ideogram according to[21]. On the bottom, its evolutionary domains: the X-added region (XAR), the X-conserved region (XCR; dotted region in proximal Xp does not appear to be part of the XCR), the pseudoautosomal region PAR1, and evolutionary strata S5–S1. (**C**) The reduced all-to-all mapping graph of HPRCy1 versus itself, with contigs represented as nodes and mappings as edges. In red contigs covering the evolutionary strata 5 and 4 on chromosome X; (**D**) Coloring the reduced homology mapping graph in C with community assignments. Panels **C** and **D** use the same layout as Fig. 1 but focus only on the X and Y region of the visualization.

## Extended Data Fig. 2 An overview of our approach to build a PVG for HPRCy1 contigs that can be anchored to a specific acrocentric q-arm.

(**A**) As input, we take the entire HPRCy1 and map it to T2T-CHM13. (**B**) This yields mappings to acrocentric chromosomes, which we filter to select contigs that map across the centromeres (red cytobands) between non-centromeric regions (over-labeled green). We include two HG002 assemblies based on standard HiFi (from HPRCy1) and on both HiFi and ONT data (from Verkko). (**C**) We then apply PGGB to build a PVG from the HPRCy1-acro collection. PGGB first obtains an all-to-all alignment of the input (C.a.), which is converted to a variation graph with SEQWISH[28] (C.b.), then normalized with sorting and multiple sequence alignment steps in SMOOTHXG (C.c-f). (**D**) The resulting PVG expresses genomes as paths, or walks, through a common sequence graph. This model thus contains all input sequences and their relative alignments to all others—in the example we see a CTGG/AAGTA block substitution between genomes 1 and 2.

## Extended Data Fig. 3 Scheme of the graph untangling.

We applied ODGI UNTANGLE to obtain a mapping from segments of all PVG paths onto T2T-CHM13. The segmentation cuts the graph into regular-sized regions whose boundaries occur at structural variant breakpoints. For each query subpath through a graph segment, we use a Jaccard metric over the sequence space of the subpaths to find the best-matching reference segment.

## Extended Data Fig. 4 Characteristics of the pseudo-homologous regions of acrocentric chromosomes on chromosome 13.

**(A)** We focus on the first 25 Mbp of chromosome 13, shown here as a red box over T2T-CHM13 cytobands. Pseudo-homologous regions (**PHRs**), where diverse sets of acrocentric chromosomes recombine, are highlighted relative to T2T-CHM13 genome annotations for repeats, GC percentage, and genes. Above, we indicate regions of interest described in the main text:

rDNA, SST1 array, centromere, and q-arm. Below, we show T2T-CHM13-relative homology mosaics for each chromosome 13 matched contig from HPRCy1-acro, with the most-similar reference chromosome at each region shown using the given colors (**Target**). **(B)** Aggregated untangle results in the SAACs. For each acrocentric chromosome, we show the count of its HPRCy1 q-arm-anchored contigs mapping itself and all other acrocentrics (**Contigs**), **(C)** as well as the regional (50kbp) untangle entropy metric (**Regional homology entropy**) computed over the contigs' T2T-CHM13-relative untanglings. **(D)** By considering the multiple untangling of each HPRCy1-acro contig, we develop a point-wise metric that captures diversity in T2T-CHM13-relative homology patterns (**Positional homology entropy**), leading to our definition of the PHRs. **(E)** The patterns of homology mosaicism suggest ongoing recombination exchange in the SAACs. A scan over T2T-CHM13 reveals that the rDNA units are enriched for PRDM9 binding motifs, and thus may host frequent double stranded breaks during meiosis. In (B-D) a gray background indicates regions with missing data due to the lack of non-T2T-CHM13 contigs. We provide the Centromeric Satellite Annotation (CenSat Annotation) track legend in Extended Data Table 1.

## Extended Data Fig. 5 Characteristics of the pseudo-homologous regions of acrocentric chromosomes on chromosome 14.

**(A)** We focus on the first 25 Mbp of chromosome 14, shown here as a red box over T2T-CHM13 cytobands. Pseudo-homologous regions (**PHRs**), where diverse sets of acrocentric chromosomes recombine, are highlighted relative to T2T-CHM13 genome annotations for repeats, GC percentage, and genes. Above, we indicate regions of interest described in the main text: rDNA, SST1 array, centromere, and q-arm. Below, we show T2T-CHM13-relative homology mosaics for each chromosome 13 matched contig from HPRCy1-acro, with the most-similar reference chromosome at each region shown using the given colors (**Target**). **(B)** Aggregated untangle results in the SAACs. For each acrocentric chromosome, we show the count of its HPRCy1 q-arm-anchored contigs mapping itself and all other acrocentrics (**Contigs**), **(C)** as well as the regional (50kbp) untangle entropy metric (**Regional homology entropy**) computed over the contigs' T2T-CHM13-relative untanglings. **(D)** By considering the multiple untangling of each

HPRCy1-acro contig, we develop a point-wise metric that captures diversity in T2T-CHM13-relative homology patterns (**Positional homology entropy**), leading to our definition of the PHRs. **(E)** The patterns of homology mosaicism suggest ongoing recombination exchange in the SAACs. A scan over T2T-CHM13 reveals that the rDNA units are enriched for PRDM9 binding motifs, and thus may host frequent double stranded breaks during meiosis. In (B-D) a gray background indicates regions with missing data due to the lack of non-T2T-CHM13 contigs. We provide the Centromeric Satellite Annotation (CenSat Annotation) track legend in Extended Data Table 1.

## Extended Data Fig. 6 Characteristics of the pseudo-homologous regions of acrocentric chromosomes on chromosome 15.

**(A)** We focus on the first 25 Mbp of chromosome 15, shown here as a red box over T2T-CHM13 cytobands. Pseudo-homologous regions (**PHRs**), where diverse sets of acrocentric chromosomes recombine, are highlighted relative to T2T-CHM13 genome annotations for repeats, GC percentage, and genes. Above, we indicate regions of interest described in the main text: rDNA, SST1 array, centromere, and q-arm. Below, we show T2T-CHM13-relative homology mosaics for each chromosome 13 matched contig from HPRCy1-acro, with the most-similar reference chromosome at each region shown using the given colors (**Target**). **(B)** Aggregated untangle results in the SAACs. For each acrocentric chromosome, we show the count of its HPRCy1 q-arm-anchored contigs mapping itself and all other acrocentrics (**Contigs**), **(C)** as well as the regional (50kbp) untangle entropy metric (**Regional homology entropy**) computed over the contigs' T2T-CHM13-relative untanglings. **(D)** By considering the multiple untangling of each HPRCy1-acro contig, we develop a point-wise metric that captures diversity in T2T-CHM13-relative homology patterns (**Positional homology entropy**), leading to our definition of the PHRs. **(E)** The patterns of homology mosaicism suggest ongoing recombination exchange in the SAACs. A scan over T2T-CHM13 reveals that the rDNA units are enriched for PRDM9 binding motifs, and thus may host frequent double stranded breaks during meiosis. In (B-D) a gray background indicates regions with missing data due to the lack of non-T2T-CHM13 contigs. We provide the Centromeric

Satellite Annotation (CenSat Annotation) track legend in Extended Data Table 1.

## Extended Data Fig. 7 Characteristics of the pseudo-homologous regions of acrocentric chromosomes on chromosome 21.

**(A)** We focus on the first 25 Mbp of chromosome 21, shown here as a red box over T2T-CHM13 cytobands. Pseudo-homologous regions (**PHRs**), where diverse sets of acrocentric chromosomes recombine, are highlighted relative to T2T-CHM13 genome annotations for repeats, GC percentage, and genes. Above, we indicate regions of interest described in the main text: rDNA, SST1 array, centromere, and q-arm. Below, we show T2T-CHM13-relative homology mosaics for each chromosome 13 matched contig from HPRCy1-acro, with the most-similar reference chromosome at each region shown using the given colors (**Target**). **(B)** Aggregated untangle results in the SAACs. For each acrocentric chromosome, we show the count of its HPRCy1 q-arm-anchored contigs mapping itself and all other acrocentrics (**Contigs**), **(C)** as well as the regional (50kbp) untangle entropy metric (**Regional homology entropy**) computed over the contigs' T2T-CHM13-relative untanglings. **(D)** By considering the multiple untangling of each HPRCy1-acro contig, we develop a point-wise metric that captures diversity in T2T-CHM13-relative homology patterns (**Positional homology entropy**), leading to our definition of the PHRs. **(E)** The patterns of homology mosaicism suggest ongoing recombination exchange in the SAACs. A scan over T2T-CHM13 reveals that the rDNA units are enriched for PRDM9 binding motifs, and thus may host frequent double stranded breaks during meiosis. In (B-D) a gray background indicates regions with missing data due to the lack of non-T2T-CHM13 contigs. We provide the Centromeric Satellite Annotation (CenSat Annotation) track legend in Extended Data Table 1.

## Extended Data Fig. 8 Characteristics of the pseudo-homologous regions of acrocentric chromosomes on chromosome 22.

**(A)** We focus on the first 25 Mbp of chromosome 22, shown here as a red box over T2T-CHM13 cytobands. Pseudo-homologous regions (**PHRs**), where diverse sets of acrocentric chromosomes recombine, are highlighted

relative to T2T-CHM13 genome annotations for repeats, GC percentage, and genes. Above, we indicate regions of interest described in the main text: rDNA, SST1 array, centromere, and q-arm. Below, we show T2T-CHM13-relative homology mosaics for each chromosome 13 matched contig from HPRCy1-acro, with the most-similar reference chromosome at each region shown using the given colors (**Target**). **(B)** Aggregated untangle results in the SAACs. For each acrocentric chromosome, we show the count of its HPRCy1 q-arm-anchored contigs mapping itself and all other acrocentrics (**Contigs**), **(C)** as well as the regional (50kbp) untangle entropy metric (**Regional homology entropy**) computed over the contigs' T2T-CHM13-relative untanglings. **(D)** By considering the multiple untangling of each HPRCy1-acro contig, we develop a point-wise metric that captures diversity in T2T-CHM13-relative homology patterns (**Positional homology entropy**), leading to our definition of the PHRs. **(E)** The patterns of homology mosaicism suggest ongoing recombination exchange in the SAACs. A scan over T2T-CHM13 reveals that the rDNA units are enriched for PRDM9 binding motifs, and thus may host frequent double stranded breaks during meiosis. In (B-D) a gray background indicates regions with missing data due to the lack of non-T2T-CHM13 contigs. We provide the Centromeric Satellite Annotation (CenSat Annotation) track legend in Extended Data Table 1.

## Extended Data Fig. 9 PRDM9 binding motif in the acrocentric chromosomes.

For each T2T-CHM13 acrocentric chromosome, we show the number of human PRDM9 binding motif hits present in windows 20 kbps long.

## Extended Data Fig. 10 Linkage disequilibrium decay with distance between markers per acrocentric chromosome.

Each LD decay plot shows the p-arm (purple), q-arm (pink), and PHR (blue) mean $r^2$ (points) and 95% confidence intervals (error bars) for marker pairs binned by the given inter-marker distance range (x-axis). Dot size is proportional to the number of pairwise comparisons within a bin. LD decay is faster in PHRs for chromosomes 13, 14, and 22. No notable LD decay is observed in PHRs for chromosome 15.

**Extended Data Table 1 Centromeric Satellite Annotation (CenSat Annotation) track legend**

## Supplementary information

### [Supplementary Information]

This file contains Supplementary Figs. 1–29.

### [Reporting Summary]

### [Peer Review File]

### [Supplementary Note 1]

Physical proximity modelling of acrocentric short arms.

### [Supplementary Files]

This file contains Supplementary Files 1–8.

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

# About this article

## Cite this article

Guarracino, A., Buonaiuto, S., de Lima, L.G. *et al.* Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335–343 (2023). https://doi.org/10.1038/s41586-023-05976-y

- Received: 15 August 2022
- Accepted: 17 March 2023
- Published: 10 May 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05976-y

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

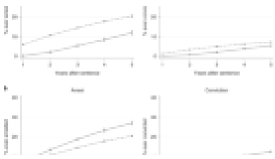
Provided by the Springer Nature SharedIt content-sharing initiative

## Human pangenome supports analysis of complex genomic regions

- Arya Massarat
- Melissa Gymrek
- Hákon Jónsson

# Nature News & Views Forum 10 May 2023

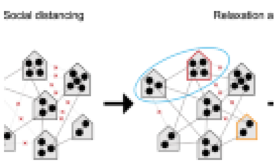
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05976-y



## A natural experiment study of the effects of imprisonment on violence in the community

13 May 2019

David J. Harding, Jeffrey D. Morenoff, … Ingrid A. Binswanger




## The effect of eviction moratoria on the transmission of SARS-CoV-2

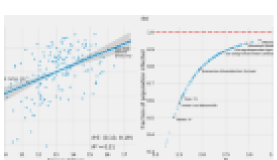
15 April 2021

Anjalika Nande, Justin Sheen, … Alison L. Hill

## Resolution of apparent paradoxes in the race-specific frequency of use-of-force by police

05 June 2018

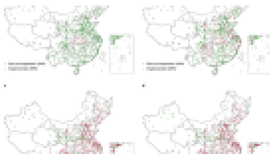
Cody T. Ross, Bruce Winterhalder & Richard McElreath



## Early pandemic COVID-19 case growth rates increase with city size

28 June 2021

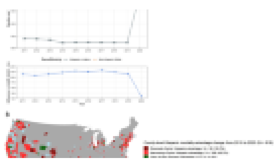
Andrew J. Stier, Marc G. Berman & Luís M. A. Bettencourt



## Short- and medium-term impacts of strict anti-contagion policies on non-COVID-19 mortality in China

29 November 2021

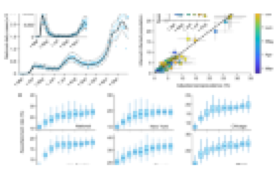
Jinlei Qi, Dandan Zhang, … Maigeng Zhou



## Trends in national and county-level Hispanic mortality in the United States, 2011–2020

12 July 2022

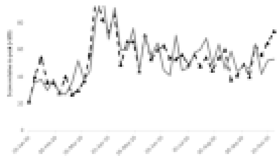
Ashish Sarraju, Summer Ngo, … Fatima Rodriguez



## Burden and characteristics of COVID-19 in the United States during 2020

26 August 2021

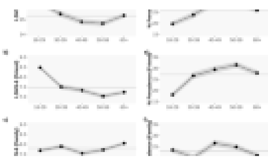
Sen Pei, Teresa K. Yamana, … Jeffrey Shaman



[COVID-19 restrictions and age-specific mental health—U.S. probability-based panel evidence](#)

04 August 2021

Elvira Sojli, Wing Wah Tham, … Michael McAleer



[Social isolation during the COVID-19 pandemic in Spain: a population study](#)

22 July 2022

Marina Martínez-Garcia, Emilio Sansano-Sansano, … Nuria Oliver

- Article
- Open Access
- Published: 19 April 2023

# COVID-19 amplified racial disparities in the US criminal legal system


- Brennan Klein ,
- C. Brandon Ogbunugafor ,
- Benjamin J. Schafer,

- Zarana Bhadricha,
- Preeti Kori,
- Jim Sheldon,
- Nitish Kaza,
- Arush Sharma,
- Emily A. Wang,
- Tina Eliassi-Rad,
- Samuel V. Scarpino &
- …
- Elizabeth Hinton

Show authors

*Nature* **volume 617**, pages 344–350 (2023)

- 7181 Accesses
- 279 Altmetric
- Metrics details

# Abstract

The criminal legal system in the USA drives an incarceration rate that is the highest on the planet, with disparities by class and race among its signature features[1,2,3]. During the first year of the coronavirus disease 2019 (COVID-19) pandemic, the number of incarcerated people in the USA decreased by at least 17%—the largest, fastest reduction in prison population in American history[4]. Here we ask how this reduction influenced the racial composition of US prisons and consider possible mechanisms for these dynamics. Using an original dataset curated from public sources on prison demographics across all 50 states and the District of Columbia, we show that incarcerated white people benefited disproportionately from the decrease in the US prison population and that the fraction of incarcerated Black and Latino people sharply increased. This pattern of increased racial disparity exists across prison systems in nearly every state and reverses a decade-long trend before 2020 and the onset of COVID-19, when the proportion of incarcerated white

people was increasing amid declining numbers of incarcerated Black people[5]. Although a variety of factors underlie these trends, we find that racial inequities in average sentence length are a major contributor. Ultimately, this study reveals how disruptions caused by COVID-19 exacerbated racial inequalities in the criminal legal system, and highlights key forces that sustain mass incarceration. To advance opportunities for data-driven social science, we publicly released the data associated with this study at Zenodo[6].

# Main

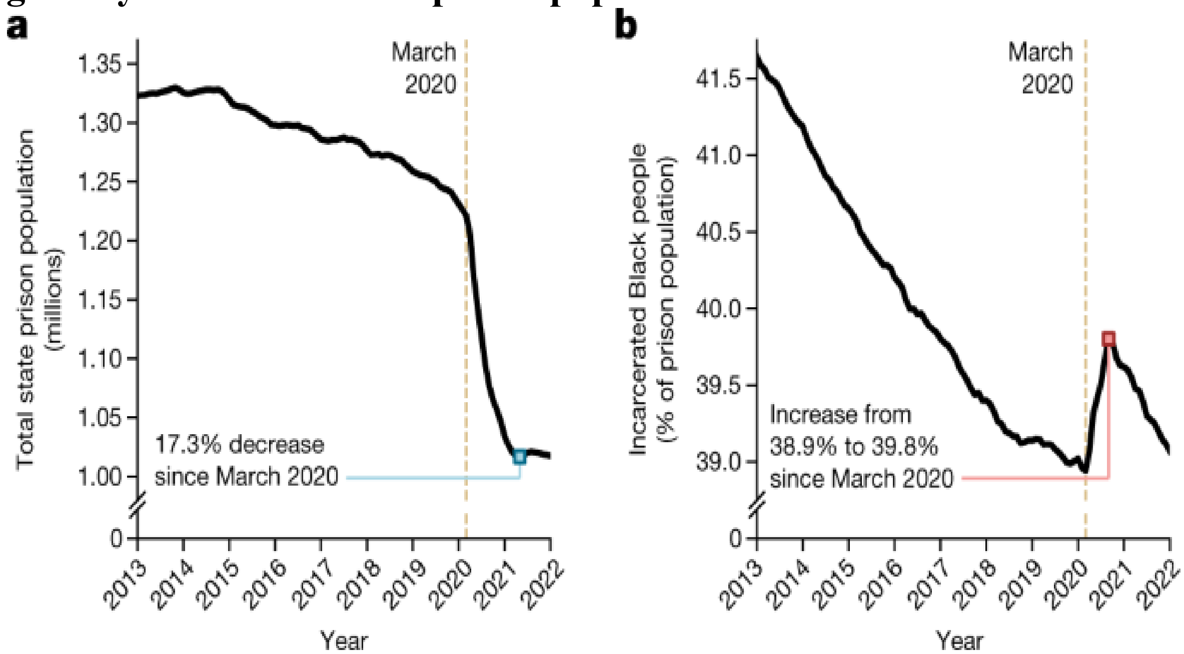
Mass incarceration in the USA is distinguished by striking racial disparities and a rate of imprisonment that surpasses that of all other nations, with 2.12 million people in prisons and jails in 2019 (refs. [1,2,3,7,8,9,10]). Owing to a combination of structural inequities and discriminatory enforcement, Black and Latino people are more likely to be stopped by police[11], held in jail pre-trial[12], charged with more serious crimes[13] and sentenced more harshly than white people[14,15]. These practices have made Black men in the USA six times as likely and Latino men 2.5 times as likely to be incarcerated as white men[16,17].

In this study, we demonstrate how the COVID-19 pandemic—which produced the largest, most rapid single-year decrease in prison population in US history—amplified existing inequities in the nation's criminal legal system[4]. Across nearly every state and federal prison system, we observe a convergent pattern: a substantial decrease in the overall number of people incarcerated (by approximately 200,000), but a meaningful increase in the proportion of incarcerated Black, Latino and other non-white people. We conclude that sentencing patterns are a central mechanism driving the racial disparity.

The trend we identify represents a substantial deviation from patterns preceding the pandemic. Before COVID-19, incarcerated Black people accounted for a declining share of the total prison population: roughly 41.6% of people incarcerated in state prisons were Black in March 2013, and by March 2020 this number had fallen to 38.9%—a decline of 2.7 percentage

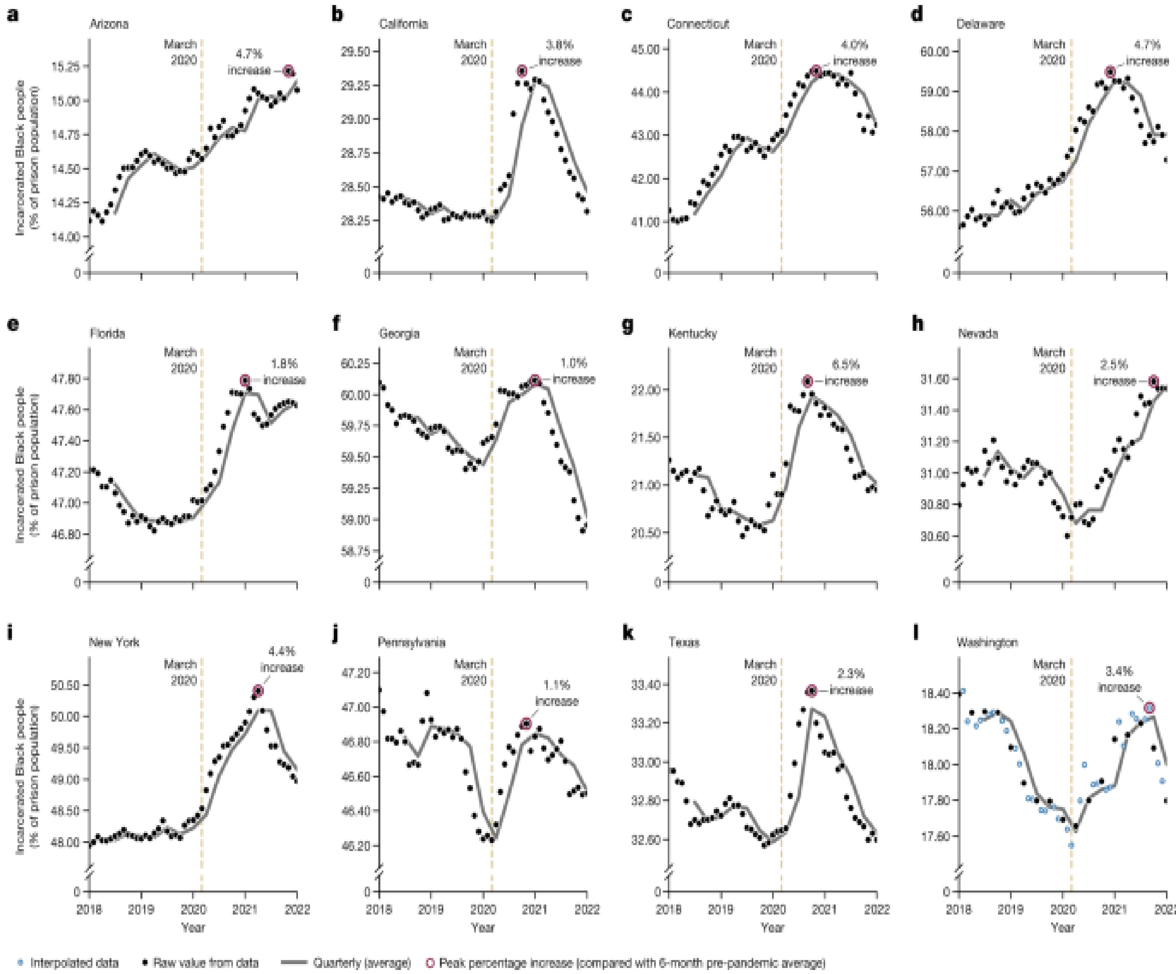
points over seven years. During the height of COVID-19 closures, from March 2020 to November 2020, this percentage increased by 0.9 points, erasing much of the progress over the last decade (Figs. 1b and 2 and Supplementary Fig. 2; see Supplementary Fig. 16 for comparison between effects among non-white versus Black populations). The trend we observe at the national level is reproduced exactly among states with the highest Black and Latino populations, and persists in some form in nearly every other state.

**Fig. 1: Dynamics of the US prison population.**




**a**, Total number of incarcerated people in the USA from January 2013 to January 2022. **b**, Total percentage of incarcerated Black people, as reported by states' Departments of Correction. According to data from the US census, Black people account for 13.4% of the total population[43]. This plot includes data from 49 states and the District of Columbia—data from Michigan are excluded as the state reports only "white" and "nonwhite" as race categories.

**Fig. 2: Percentage of incarcerated Black people across 12 representative states.**




In some states, the percentage of incarcerated Black people had been decreasing over the past several years. In others, this percentage had been increasing. Across a variety of pre-2020 trends, we see the same general trend across the USA: during the pandemic, incarcerated Black people accounted for an even larger share of the total prison population than in previous years. **a**–**l**, We plot this trend in Arizona (**a**), California (**b**), Connecticut (**c**), Delaware (**d**), Florida (**e**), Georgia (**f**), Kentucky (**g**), Nevada (**h**), New York (**i**), Pennsylvania (**j**), Texas (**k**) and Washington (**l**)—states with consistent, frequent data reporting. Although each panel has a different vertical scale, the inset text indicates the peak percentage increase relative to 6-month pre-pandemic averages for each state.

Data reporting methods on racial demographics in prisons have made it difficult for researchers to disentangle the various mechanisms driving observed disparities in incarcerated populations. We manually assembled and validated a dataset covering all 50 US states, the District of Columbia

and the Federal Bureau of Prisons to both quantify the widening racial disparity observed during the first year of the COVID-19 pandemic and uncover its plausible causes. The result of this newly assembled, public dataset—comprising over 9,000 records across more than 20 years—is a view into the dynamics of prison populations before, during and after the onset of the pandemic. We publicly released the data associated with this study at Zenodo[6].

Overall, the number of incarcerated people decreased markedly in 2020. However, we show that the magnitudes of these declines were not equal by race, especially for incarcerated Black people (Fig. 3). We estimate that nearly 15,000 fewer Black people would have been incarcerated in January 2021 if the racial disparities we observe were not present (Supplementary Table 6). We discuss this observed disparity and related observations in light of the ethics of public health interventions, national debates about the future direction of policing and incarceration, and the importance of data infrastructure in responsible public policy. These discussions highlight how sentencing and other policies that seem to be 'race blind' can nonetheless lead to outcomes that are skewed by race[18]. We speculate that our findings transcend the influence of COVID-19, and discuss how large-scale disruptions can have a clear, quantifiable signature on extant inequalities.

## Declining incarcerated populations

The population of people incarcerated in US state prisons decreased by at least 17% between March 2020 and July 2021, from approximately 1.23 million to 1.02 million (Fig. 1a). A decrease in prison population occurred in every state, and, in most, started in early to mid-April 2020 (see Supplementary Fig. 1 for a state-by-state look at prison populations over time). This nationwide trend persisted despite stark differences in state-level trends pre-2020: For instance, many states entered 2020 with a steadily declining prison population (for example, total incarcerated population in Massachusetts fell from 11,403 in January 2013 to 8,292 in January 2020—declining 4.4% per year on average; in South Carolina: 22,146 in January 2013 to 18,041 in January 2020, an average of 2.9% decrease per year; in California: 134,534 in January 2013 to 124,027 in January 2020, an average of 1.1% decrease per year; and so on), others had relatively stable prison

populations (for example, Virginia: 29,875 in 2013, 29,233 in 2020; Georgia: 56,951 in 2013, 55,218 in 2020) and many had growing prison populations before COVID-19 (for example, in Idaho, there were 8,030 people incarcerated in January 2013 and 9,502 in January 2020, increasing on average 2.5% per year; in Montana: 2,452 in January 2013 to 2,806 in January 2020, increases of 2.0% per year, on average; Supplementary Fig. 1). Nevertheless, we see large reductions in the overall prison population across every state in the USA during the pandemic, with magnitudes ranging from a 5.8% reduction in Nebraska to 37.2% in New Jersey, and the median state's prison population falling by 18% of pre-pandemic levels. In Supplementary Table 2, we detail the scale and timing of each state's population decline.

As of January 2022, several states' prison populations continued to decrease (Arizona, Massachusetts, Washington, Louisiana and Pennsylvania, among others) and did so steadily throughout the pandemic. Other states' prison populations dropped sharply in the early months of the pandemic but saw their prison populations begin to increase again by January 2021 (Montana, Idaho, Iowa, Utah, Nebraska, West Virginia and California, among others) or by July 2021 (Wisconsin, Ohio, Texas, Connecticut, Rhode Island, Kentucky and so on). In Supplementary Fig. 1, we plot time series for each state's prison population over the past several years. Additionally, in Supplementary Table 1, we give an overview for each state's approach for reporting prison population statistics, along with how we collected each state's data.

# Changing racial demographics in prisons

Despite an overall decline in the total incarcerated population during the pandemic, there was an increase in the proportion of incarcerated Black people (Fig. 1b). This increase in racial disparity occurred nationally and in nearly every state, transcending vast differences in approach to crime and incarceration. In Fig. 2, we show the percentage of incarcerated Black people across 12 states (see Supplementary Fig. 2 for these trends in every state and the Federal Bureau of Prisons). However, the spike in the proportion of Black people in prison was temporary in most states, eventually returning to pre-pandemic levels by the end of 2021. We explore

possible explanations for this reversal in subsequent sections, but the most likely reason is that the pace of prison admissions—which typically have a lower Black–white racial disparity than the overall incarcerated population[19]—began to approach pre-pandemic rates in early 2021.

Although the national trend we identify in Fig. 1 (that is, an abrupt increase in the proportion of Black people incarcerated) occurred in most state-level prison systems, there were meaningful differences that suggest possible mechanisms behind the disparity (Fig. 2 and Supplementary Figs. 2–4). In Fig. 2, we highlight several examples of state-level variability in the proportion of Black people incarcerated; for instance, trends in states such as Georgia, Kentucky and Texas resemble the shape seen nationally, whereas states such as Connecticut and Delaware saw an already-increasing trend in the percentage of incarcerated Black people increase even faster after March 2020. Five states—Maine, Maryland, Missouri, Oregon and Wyoming—are the only prison systems in the USA that do not clearly conform to the pattern we see across the country (a few states in Supplementary Fig. 2—for example, Missouri and Oklahoma—technically fit our criteria for exhibiting this trend but only weakly). These five states have a combined incarcerated population that amounts to roughly 5% of the national total and offer important insights into the underlying mechanisms behind the trends we see nationally. Namely, each of these states has either a relatively small proportion of incarcerated Black people or a prison system with fewer people with shorter-term (for example, fewer than 2 years) sentences compared to nationwide averages. We will show that the latter is probably the more powerful force contributing to the overall nationwide trend.

Ultimately, these observations lead us to outline three explanations that could bring about the trends from Fig. 1: who is admitted to prison; who is released from prison; and who remains in prison. These proposed mechanisms demonstrate different levers through which the pandemic may have influenced the racial composition of the incarcerated population, and dovetail with existing research on the dynamics of the American carceral state.

# Mechanisms of disparity

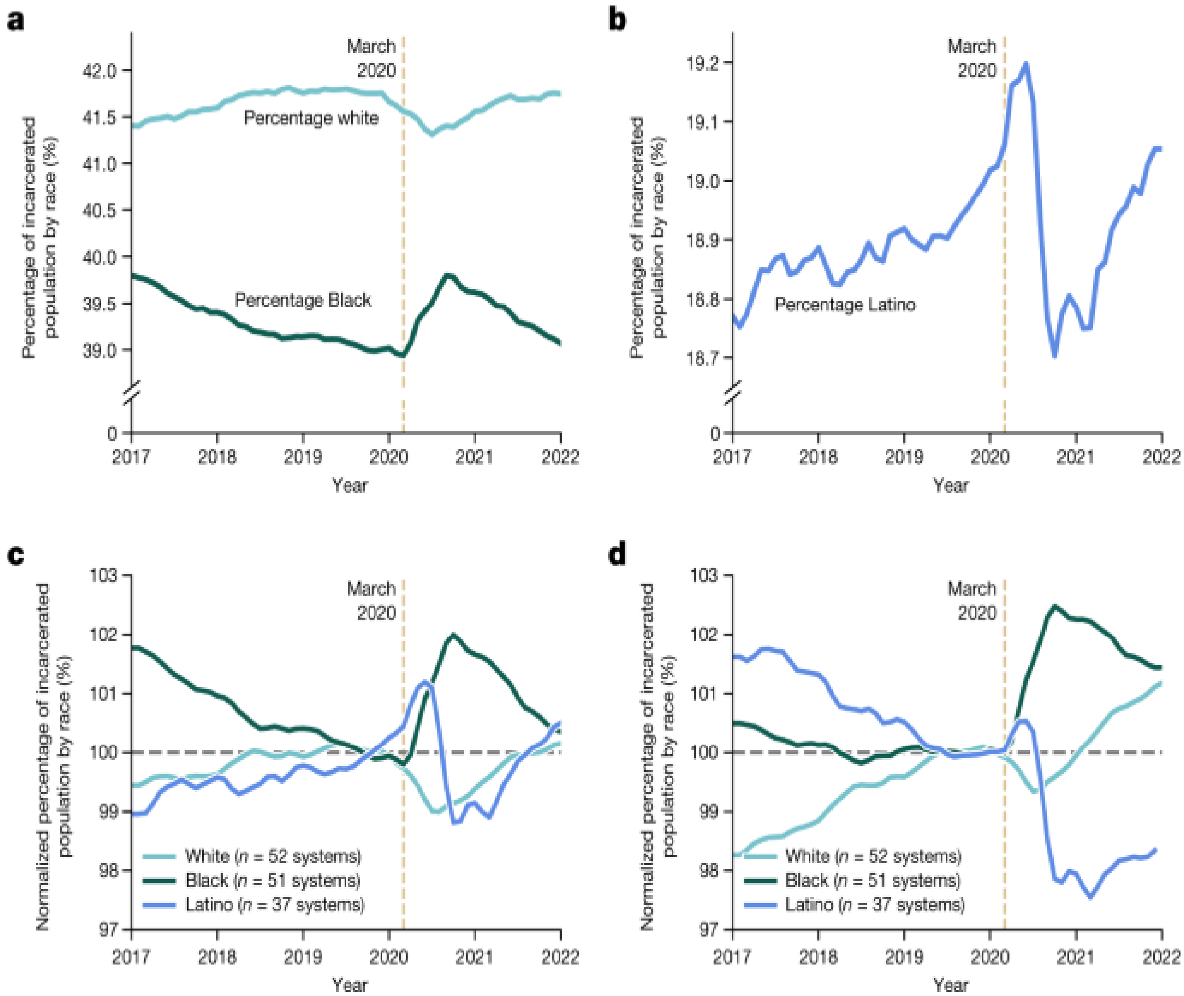
Consider a time series of a state's prison population that does not notably change over several years. For this to occur, there needs to be approximately the same number of admissions and releases. For the demographic makeup of the prison population to remain stable, the relative number of admissions and releases by race also needs to be roughly equivalent over time. If there are sustained periods with more admissions (or releases) of a certain demographic, that will skew the overall distribution of the prison population.

Understanding the dynamics of admissions, releases and sentencing offers us a path towards identifying and isolating potential mechanisms that could bring about a steadily declining rate of incarceration of Black individuals (seen for nearly a decade before the pandemic), and the subsequent spike in the proportion of incarcerated Black people during the COVID-19 pandemic. Namely, the observed spike in Fig. 1b must be due to a disparity in who was admitted to prison during the pandemic, who was released or a combination of both.

## Admissions

In every state except Nebraska, courts closed at the beginning of the pandemic. These closures substantially reduced or altogether halted admissions into prisons for several months, starting around April 2020 (refs. [20,21,22,23]). The Virginia Department of Corrections acknowledges the causal effect of court closures on the state's incarcerated population in their 2020 Annual Report[24], "The reduction in [average daily population] is directly attributed to the suspension of intake due to COVID-19". Similarly, a spokesperson for the Michigan Department of Corrections estimated that half of the reductions in incarcerated population were due to a decline in new admissions from courts and county jails[25] .

**Fig. 3: Comparison of Black, white and Latino incarcerated populations over time.**

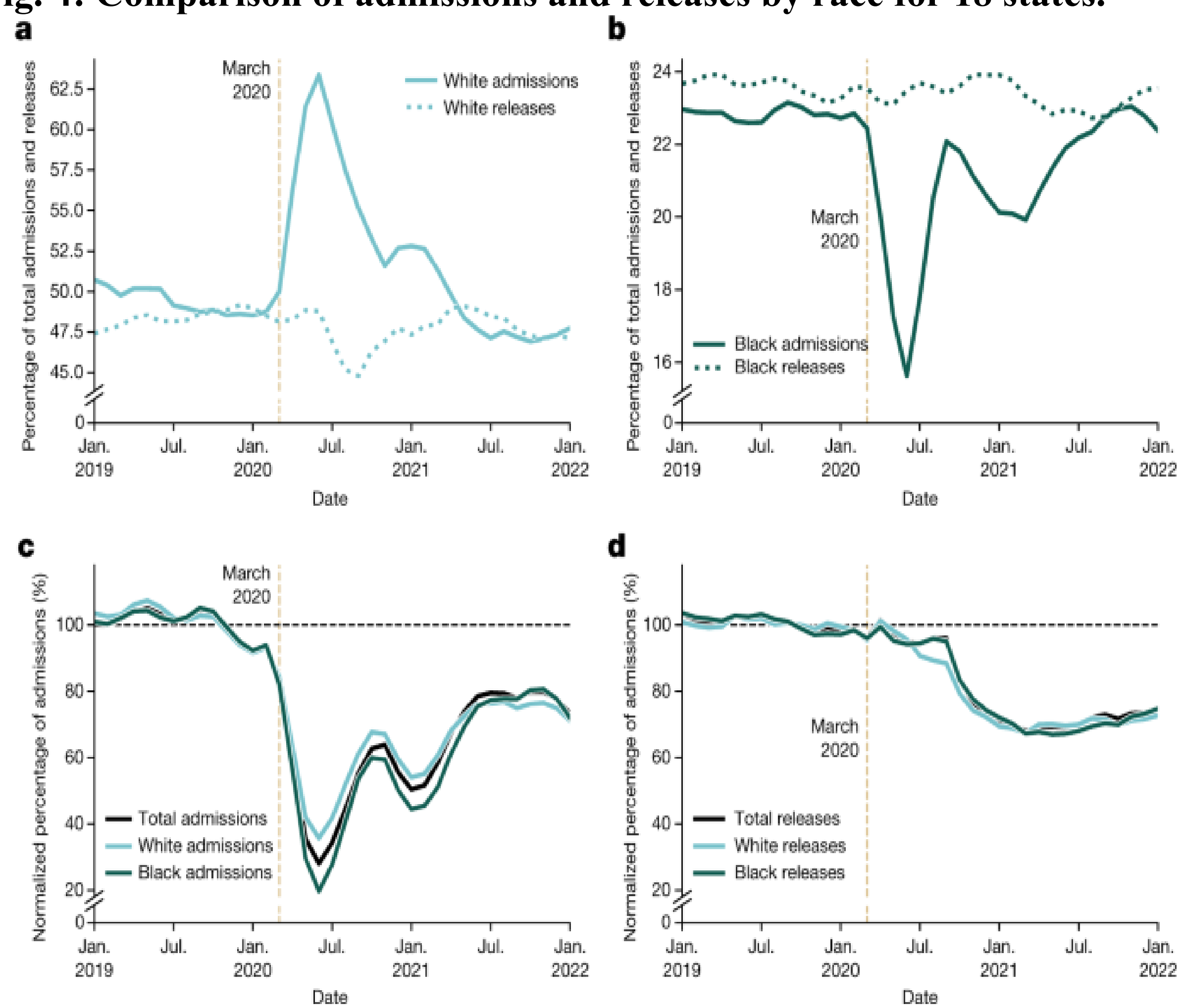



**a**, Percentage of incarcerated population who are Black and white. Note especially that the effect size of the demographic changes during the COVID-19 pandemic are more pronounced in the incarcerated Black population (see Supplementary Fig. 16 for additional comparisons). **b**, Percentage of incarcerated Latino people. **c**, Percentage of white, Black and Latino incarcerated populations, normalized by the average value between March 2019 and February 2020. **d**, Percentage of white, Black and Latino incarcerated populations, normalized by the slope of each curve between March 2019 and February 2020.

On the basis of admissions data from 18 states, we estimate that the total monthly admissions to prison fell to about 30% of pre-pandemic averages by May or June 2020 (Fig. 4). This reduction in admissions provides a potential mechanism behind the sharp increase in the percentage of incarcerated Black people in Fig. 1b. Specifically, systematic racial differences occurring in

monthly prison admissions during this period could drive changing disparity in the demographics of the incarcerated population. However, data from the 18 states presented in Fig. 4c actually show the reverse (that is, the percentage of Black individuals admitted to prison fell even lower than the corresponding rate for admissions of white individuals). Although we do see abrupt spikes in the percentage of Black individuals admitted to prison in a few states (for example, Wisconsin and Texas), this proposed mechanism seems not to be widespread enough to explain the nationwide trends we observe.

**Fig. 4: Comparison of admissions and releases by race for 18 states.**




Although data on monthly admissions to and releases from prison are less readily available than prison population data, we can nevertheless highlight the average dynamics of 18 states' data. **a**, Percentage of total monthly admissions (solid) and releases (dotted) for individuals who are white. **b**,

Percentage of total monthly admissions (solid) and releases (dotted) for individuals who are Black. **c**, Normalized comparison of the change in monthly white, Black and total admissions. **d**, Normalized comparison of the change in monthly white, Black and total releases. Note: the vertical scales in **a** and **b** are different, whereas **c** and **d** share the same vertical scale.

Data from Florida offer another example of how changes to court proceedings influence prison admissions and the racial distribution of people admitted to prison. In Supplementary Fig. 18 we plot monthly trial statistics from circuit criminal defendants in Florida; after March 2020, we see sharp declines in the number of disposed defendants in Florida Circuit Criminal Courts, as well as the percentage of filed defendants that become disposed. Amid these declines, we see an abrupt increase in the percentage of cases that were dismissed before trial (that is, defendants whose charges were dropped). In Supplementary Fig. 19, we report that an increased proportion of the defendants with pre-trial case dismissals were white in the months after the start of the pandemic.

Prison admissions may also decline owing to policy changes or disruptions to a common source of prison admissions: county jails. Although there continues to be poor standards for reporting and maintaining these kinds of data, this potential source of prison admissions is important for a nationwide story of mass incarceration during the COVID-19 pandemic. Despite variability in admissions playing a role in the racial distribution of incarcerated populations, changing disparity in admissions alone does not seem to be widespread enough to account for the nationwide trends we observe in Fig. 1b. As we will see in the following section, a similar story emerges when looking at the demographics of people released from prison.

## Releases

In an effort to reduce the risk of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission, several states enacted policies designed to de-densify prisons. Depending on the state, these directives came from executive orders from the governor, state legislatures or governing boards. In Utah, for example, policies around releases are designed, approved and implemented by the Board of Pardons and Parole—an entirely separate entity from the courts and the Department of

Corrections. According to the Board of Pardons and Parole, incarcerated people who are eligible for early release needed to be already characterized as a non-violent offender, be within 90 days of release (this was later extended to 180 days[26]) and have an approved address to stay at after their release.

In Arkansas, an authorization from Governor Hutchinson (Executive Orders 20-06 and 20-16 (ref. [27])) made 1,243 incarcerated people eligible for early release as of 30 April 2020. Those deemed eligible needed to have a parole plan in place, be medically screened (that is, tested and screened for symptoms of COVID-19) and undergo final approval by the Arkansas Department of Corrections director to be released. In Supplementary Section 4.3 and Supplementary Fig. 23, we show that disproportionately more white people were released in Arkansas through this effort. The racial disparity in who was released by Governor Hutchinson's orders is due to the overlap between the state's release eligibility criteria and the racial differences in sentence classification—a tension we discuss further in the section in the Methods entitled Release policy data.

In Fig. 4d, we plot estimates of the nationwide change in monthly releases as a percentage of pre-pandemic values. What we see is that, despite efforts to reduce prison density through targeted releases, the rate of prisoner release was lower during much of the pandemic. At its lowest value (between February and May 2021), the number of people released from prison each month reached nearly 70% of pre-pandemic values. In the absence of changing admission patterns, this decline in releases should have led to an increase in the total incarcerated population in the USA, which is the opposite of the pattern we see in Fig. 1a. Therefore, we can conclude rather strongly that changing release rates did not drive the reduction in the incarcerated population during the pandemic. We also do not find meaningful differences in the relative number of releases by race during this time period; if anything, these data suggest that during the early months of the COVID-19 pandemic, Black people accounted for a higher percentage of monthly releases compared to pre-pandemic averages. As was the case for demographics of prison admissions, disparities in the monthly releases are unlikely to be driving the trends in Fig. 1b. However, data on prison releases point to an important, related process underlying the demographic patterns

in incarcerated populations; in the next section, we focus less on those released from prison in any given month, but rather, those who remain.

## Sentencing

On the basis of demographic data from 18 state prison systems, racial disparities in admissions and releases alone are not able to explain the broad trends observed nationally (Fig. 4). In fact, if these were the only factors influencing prison population demographics, we would expect the opposite effect seen in Fig. 1b, as we observe a large increase in the proportion of admissions of white individuals after the start of the pandemic, amid large decreases in the proportion of admissions of Black individuals and relatively commensurate rates of releases. There are examples of individual states that show sudden increases in the relative amount of Black people admitted to prison at the start of the pandemic (see Texas, for example, in Supplementary Fig. 20). Similarly, there are examples of large-scale releases causing an abrupt increase in the percentage of incarcerated Black people (see a recent example in January 2022 in data from the Federal Bureau of Prisons; Supplementary Fig. 9). Nevertheless, we do not see these factors as being anywhere near as influential as disparities in sentencing of people already incarcerated at the start of the pandemic.

In short, the most important factor underlying the dynamics in Fig. 1b is related to differences in the average sentence length of incarcerated people by race. As a statistical observation this point is quite simple: provided there are differences in the average length of prison sentence by race (for example, the average incarcerated Black person serving a longer prison sentence than the average white incarcerated person; see Texas as an example in Supplementary Fig. 22) and sustained reductions in new admissions (as in Fig. 4c), then we will expect to see the effect observed in Fig. 1b. In addition to that basic mechanism, one can imagine factors that would exacerbate and/or attenuate the size and timing of the spike. These include: new or atypical patterns in prison admissions by race (relative to averages before the decline in admissions), or new or atypical changes in prison releases by race.

By casting sentencing differences as the driver behind the observations in this study, we are able to better understand why the main effect in Fig. 1b is

so pronounced among incarcerated Black people and less so (although still present) when looking at incarcerated Latino people (Fig. 3). Illinois and Texas offer two particularly powerful examples that show increases in their proportion of incarcerated Black people. As one would expect given our proposed mechanism, the median sentence length for incarcerated Black people in each of these states is higher than that of white people. However, when we compare the median sentence lengths between incarcerated Black and Latino people, we find high Black–Latino overlap in Illinois but high white–Latino overlap in Texas. That is, white people in Illinois serve shorter sentences on average than Black and Latino people, but in Texas, white and Latino people serve shorter sentences than incarcerated Black people. According to the mechanism proposed above, we would expect this baseline difference in sentencing lengths to produce pandemic-related spikes in the percentage of Black and Latino people in Illinois and produce spikes only in the percentage of incarcerated Black people in Texas. This, in fact, is what we observe (Supplementary Fig. 7).

The policies, societal disruptions and behavioural changes that emerged following the onset of the COVID-19 pandemic amplified existing and long-standing racial disparities in the US carceral system. Consistent with other research, our findings show that disparities in sentencing by race are core to maintaining structural inequalities in incarcerated populations[14,19].

# Discussion

After declining steadily for the past decade, the percentage of Black and other non-white incarcerated people increased sharply during 2020, a trend that was present in almost every prison system across the country. To identify the mechanisms behind this increasing racial disparity, we collected and validated a dataset that includes state-level information on police encounters, court proceedings and incarcerated populations. To obtain such granular information across all 50 states, the District of Columbia, and the Federal Bureau of Prisons, we manually collected data from individual Departments of Corrections and filed numerous Freedom of Information Act requests (Methods).

Before the COVID-19 pandemic, racial disparities in admissions were smaller than disparities in the prison population, as recent trends show a migration towards a class-driven disparity, with lower-educated white people steadily increasing in their rate of admission[28]. In a sense, courts had been serving as an instrument for decreasing the racial disparity in prisons before the pandemic; for example, the Black–white disparity in prison admissions is typically a ratio of 2:1, whereas it is closer to 6:1 for the total incarcerated population[19] (see ref. [29] for a recent exploration of several factors underlying this trend). Thus, when court proceedings or transfers from county jails are disrupted (that is, admissions are reduced), the racial disproportionality in the total prison population accelerates, as observed in Figs. 1 and 2 and Supplementary Figs. 1 and 2. These dynamics, happening within prison systems nationwide, that see incarcerated Black and other non-white people sentenced for longer periods of time on average[14], led to the abrupt nationwide increase in the percentage of incarcerated Black people, starting in March 2020. Differences in the length of sentence by race seem to be a key factor in producing the trends from Fig. 1b, but this effect will then be compounded if—in addition to overall decreases in admissions—there are also sudden changes in the typical distribution of the race of people admitted into prisons, which we see, for example, in Texas during the summer of 2020 (Supplementary Fig. 20b).

Understanding the role that racial disparities in sentencing play in producing the trends from Fig. 1b is key for making predictions about how sudden societal disruptions or policy changes in the future may impact prison population demographics (for example, continued pandemic, Supreme Court decisions, widespread social protests and so on). These findings can, in turn, help inform policy reform efforts. The sentencing disparity mechanism described in this work is even useful for explaining the dynamics behind the five states that did not conform to the overall national trend in Fig. 1b (Maine, Maryland, Missouri, Oregon and Wyoming; Supplementary Fig. 4). These five states maintain prison systems that incarcerate, on average, fewer people under shorter (less than 2-year) sentences, according to data from the National Corrections Reporting Program[30] (differences explored in Supplementary Figs. 5 and 6). This observation that states with fewer short-term prison sentences did not show the same racial disparity we found nationally has two subtle but important consequences. First, it suggests that

a key reason why the disparities emerge is due to releases of incarcerated people who served shorter-term sentences (without a corresponding amount of admissions). This makes sense, because on any given day, a randomly selected person being released from prison is likely to have been sentenced for a shorter time period. Second, if white people are more likely to serve shorter sentences, then an overall reduction in the amount of people serving shorter-term prison sentences means there are fewer people serving shorter-term sentences who could be 'eligible' to drive the main effect in Fig. 1.

Although racial disparities in sentence lengths seem to be the most robust explanation behind the trend in Fig. 1b, we want to avoid disregarding the potential effects that racial disparities in prison admissions could have played during the COVID-19 pandemic. For example, there is another well-known mechanism through which court closures could have affected different states' relative rates of Black and Latino prison admissions during the pandemic (Supplementary Fig. 18): relative increases in pre-trial case dismissals (Supplementary Fig. 18d) and pre-trial plea deals. Plea deals in particular have long been demonstrated to result in a disproportionate number of Black defendants spending time in prison[13,31,32]. Interruptions in court proceedings may have contributed to the increased Black and Latino representation in prison populations by: reducing the increasingly large flux of admissions of white individuals to prison; amplifying processes—pre-trial case dismissals and pre-trial plea deals—that are long understood to be a leading contributor to disparities in judicial outcomes for Black individuals. Disruptions in the typical, pre-pandemic court proceedings also offer a compelling explanation as to why (as seen in Fig. 1b) we see the reversion to pre-2020 levels, starting in early 2021: the reduction in admissions stopped and, in most states, the total incarcerated population began to increase once again (Fig. 2 and Supplementary Fig. 1).

Beyond disparities in sentencing and admissions, the COVID-19 pandemic provided several specific challenges that shaped release patterns. Maintaining the largest and most expansive prison system in the world is a major challenge to public health, especially in the context of infectious diseases[4,33]. In particular, severely overcrowded conditions have presented a public health threat during the COVID-19 pandemic[34]. The physical and administrative structure of prisons provided constraints on ways to

quarantine incarcerated people and de-densify congregate settings[34,35,36,37,38]. In recognition of these circumstances, several states enacted policies and initiated executive orders to release individuals who they deem eligible[23]. As a public health intervention, decarceration is a highly effective way to mitigate outbreaks inside and outside prisons[4,34,36,37,38,39,40,41]. During the pandemic, criteria for decarceration differed from state to state, but often included factors such as the age of the incarcerated person and the offence for which they were convicted (for example, non-violent drug offenders)[42]. We were able to quantify disparities in some states' efforts to de-densify prisons (for example, in Arkansas; see Supplementary Section 4.3 and Supplementary Fig. 23), which suggests that even decarceration policies widely understood to be consistent with effective and ethical public health practice (and that are assumed to be 'race blind') are susceptible to existing structural and racial inequalities. Moreover, one of the most important consequences of disparities in releases is not only about who is released, but who is left behind: the increase in the proportion of incarcerated Black and other non-white people translates to their being at a heightened risk of exposure to SARS-CoV-2.

Taken together, our findings reveal that the pandemic provided a 'stress test' for the criminal legal system. In engineering, stress tests involve exposing an apparatus to extreme conditions to reveal its fragilities; under these conditions, it can be easier to uncover the mechanisms that govern it. Using a range of data sources, we have argued that COVID-19 amplified underlying racial disparities in the carceral state. As is the case with many complex systems, the dynamics of prison populations are defined by interactions between multiple actors that, in combination, create unexpected or troubling results. In response to these findings, society has an ethical obligation to act, and reform sentencing practices and the broader criminal legal system towards more equitable ends.

# Methods

## State prison populations over time

Time series data about states' prison populations over time were collected manually through scraping Departments of Corrections websites, as well as

direct requests to state officials through public record requests (for example, Freedom of Information Act requests and so on). For every state in our dataset, we sought the most temporally resolved data as possible. We collected population data at either weekly, monthly, quarterly or, for some states, yearly levels. The most common form of data we were able to collect is the number of people incarcerated at a given time in a given state, on a monthly timescale. In Supplementary Table 1, we link to the data source for every state in our dataset, and in Supplementary Section 1, we show how the prison population of every state has changed over time.

We compared the data collected here to data from other organizations that report statistics about the US prison population—the Bureau of Justice Statistics and the Vera Institute for Justice[44]—and find high overlap between all three of the datasets. In Supplementary Section 3, we identify areas in which our data differ from those of the Bureau of Justice Statistics, and we offer an explanation for why we are confident in accuracy of our approach (for example, in several cases, we received the data directly from the states' Departments of Corrections, through public records requests).

For every state in this dataset, the total prison population includes both male and female incarcerated people (something that is not always the case in studies about the US carceral system, which so often focuses on male incarcerated people). In Alaska, New Mexico, Vermont and California, “Transgender”, “Other” or “Non-Binary” are also listed as gender categories, although this practice is not widely adopted in reporting statistics about the incarcerated population. In 27 states, incarcerated race statistics are separated by “male”, “female” and “total”, and further characterizing the interaction between race and sex in biases in admissions and releases during the COVID-19 pandemic remains future work.

## State policy data

### Court closures and reduced admissions

Qualitative data on the closure and reopening of all 50 state court systems were collected primarily through the administrative orders and/or press releases of each state system's Supreme or Superior Court or chief judicial

officer as well as through local news coverage. Most states suspended all in-person proceedings with the exception of limited emergency matters between 12 March and 20 March 2020. Several states that adopted policies early in this period issued increasingly strict guidance as the pandemic worsened. New Jersey, for example, suspended new trials on 12 March and issued a 2-week suspension on municipal court proceedings on 14 March before finally suspending all proceedings (with emergency exceptions) on 15 March. In addition to closing judicial buildings and suspending proceedings, most court closures also extended statute of limitations and filing deadlines owing to pandemic disruption. A handful of states, Pennsylvania and Texas among them, permitted or encouraged courts to begin conducting remote proceedings in their initial closure orders, although the adoption of remote proceedings was not widespread in this initial lockdown stage.

Court reopening policies were more heterogeneous than the initial closures, although trials remained suspended in most states through at least early summer 2020 (and in most cases substantially later). The earliest such policies appeared at the beginning of April 2020, with most aimed at giving regional and local judges discretion to begin hearing proceedings remotely (for example, Louisiana, Massachusetts, Florida and Iowa, among others). A substantially larger group of states adopted reopening guidelines between late April and mid May, many of which allowed essential judicial staff to return to offices following new public health guidance while also maintaining remote proceedings and expanding the number of non-trial proceedings that courts could conduct remotely. Further reopenings and the resumption of limited in-person proceedings took place in many states throughout June, July and August 2020, although trial proceedings remained suspended. Notably, several states, especially those that adopted phased reopening plans, restricted in-person proceedings and further delayed trial resumption with the autumn–winter 2020–2021 COVID-19 surge. In many states, most administrative orders restricting court operations have at the time of publishing been rescinded, although others, California notably among them, still retain certain accommodations including the option for remote proceedings.

**Release policy data**

Data on COVID-19 release policies, when they existed, were collected from states' individual corrections and prison bureau systems, governors' executive orders and local news coverage. Fifteen states did not adopt any official release policy, although our data nevertheless show that there were still reductions in the overall prison population during the pandemic in all of these states. The remaining 35 states adopted policies with varying degrees of specificity and effectiveness, although many overlapped in their broadest contours, allowing consideration for early release to be granted to incarcerated people at increased public health risk (either due to age or underlying health condition) and for those nearing parole and/or the end of their prison sentences.

Almost all states with such policies did, however, adopt a restriction preventing the release of those incarcerated for violent crimes or sex offences. North Dakota was an outlier in this regard. Of the 120 people the state initially released from prison in March 2020, 14 were serving time for violent crime convictions and 11 were convicted of sex offences. New York's release policy was notably more restrictive (on paper at least) than that of many other states—only those incarcerated for "non-criminal technical parole violations" were eligible for COVID release. As an example of one state's release policy, we include below an excerpt from the Virginia Department of Corrections' policy on releases[45], from 24 April 2020.

"The Director of the Department of Corrections is authorized to consider early release for individuals with less than one year left to serve while the COVID-19 emergency declaration is in effect. Offenders convicted of a Class 1 felony or a sexually violent offense are not eligible for consideration. The exact number of individuals eligible for early release consideration will change depending on the length of the emergency declaration order. The [Department of Corrections] will identify those that are eligible for consideration using the procedures it has developed to ensure public safety and will notify offenders who are to be released under the early release plan. A diagnosis of COVID-19 is not a release factor.

The following Early Release Criteria will be utilized in considering an incarcerated person for early release pursuant to legislation:

- Release Date: The inmate's Good Time Release Date must be calculated and verified in order for the incarcerated to be considered.

- Inmate Medical Condition: The inmate's medical condition will be considered.

- Offense History: By legislative mandate, early release does not apply to inmates convicted of a Class 1 felony or a sexually violent offense. Consideration for early release will be based on the seriousness of the current offense, in descending order as follows: Non-violent Offense, Felony Weapons Offenses, Involuntary Manslaughter, Voluntary Manslaughter, Robbery, Felony Assault, Abduction, Murder, Sex Offense.

- Viable Home Plan: The incarcerated person must have a documented approved home plan to be considered.

- Good Time Earning Level: The inmate's current good time earning level must be I or II to be considered.

- No Active Detainers: Inmates must have no active detainer to be considered.

- No Sexually Violent Predator Predicate Offenses: Inmates convicted of one or more sexually violent offenses established in §37.2-903 of the Code of Virginia are not eligible pursuant to legislation.

- Recidivism Risk: Inmates must have a risk of recidivism of medium (5-7) or low (1-4), as identified by the validated COMPAS instrument, to be considered."

Note especially the inclusion of the COMPAS risk assessment tool, which is used in court systems across the USA as a way of quantifying an offender's likelihood of reoffending (recidivism). Over the past several years, we have seen a growing body of scholarly work devoted to identifying problematic and harmful racial and economic biases that arise when algorithmic risk assessment tools are used in practice[46,47,48,49,50,51]. COMPAS, in particular, has been the subject of a number of studies that take a critical look at the

effectiveness—and ethics—of these risk assessment tools in the justice system[47,52]; in one study, COMPAS was found to predict recidivism 61% of the time, but at the same time, Black people were almost twice as likely to be labelled as high risk for reoffending but not actually reoffend[52].

Further research is needed to quantify demographic patterns in the incarcerated individuals who were released across different states, and because there was such high heterogeneity in different states' policies, it remains an open question whether we will see the same broad, systematic racial differences among the people who were released. However, as has been the case throughout the COVID-19 pandemic, the heterogeneity of policy responses across localities has typically had detrimental effects on our collective response to the pandemic[53].

## Study definitions of race and ethnicity

The data that we collected for the study used definitions of racial and ethnic groups that were determined by the agencies that collected the data. When the authors are discussing race and ethnicity in their interpretations, they are referring to the historical categories that have social, cultural and political consequences. We use the term Latino to describe people who are otherwise described as Hispanic in many settings. We have used the term non-white in select locations, as not all states had data disaggregated into the same set of categories. Thus, for some analyses, the term non-white directly describes the available data. For a list of the race categories reported by every state in our dataset, see Supplementary Table 7.

Recent advances in medical conventions have prompted discipline-wide introspection about the ways that race and ethnicity are discussed and used in research[54]. This is of critical importance to health equity and racial justice, and although in this work we rely on race statistics reported by states' Departments of Correction, future work will critically examine the differences in approaches for reporting race and ethnicity statistics of incarcerated populations. Notably, it is important to know whether a state's statistical reports use race categories that have been self-reported by the incarcerated person or whether it is interviewer-observed, which is often the case in administrative databases. These approaches are quite different and

often result in inaccuracies in measurement of racial disparities[55]. Last, in Supplementary Section 3.3, we introduce a dataset that contains policies from 48 states and the Federal Bureau of Prisons about whether race data of incarcerated individuals are obtained through self-report or visual-assignment from administrators or staff.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The incarceration data used in this work are public records in each state, and we have included the source URLs in Supplementary Table 1. Together, data from all 50 states, the District of Columbia and the Federal Bureau of Prisons create the The Dataset on Incarcerated Populations, which we have made publicly available at Zenodo[6] and at GitHub (https://github.com/jkbren/incarcerated-populations-data). The source data used to construct the Dataset on Incarcerated Populations are available through direct download using the links provided in Supplementary Table 1, by public records request or by request to the corresponding author(s).

## Code availability

The Python code to reproduce the analyses and construction of the database is available at GitHub (https://github.com/jkbren/incarcerated-populations-data) and at Zenodo[6]; these repositories contain several Jupyter notebooks with analyses and tutorials on how to automate the collection of some of the data used here.

## References


1. Muhammad, K. G., Western, B., Negussie, Y. & Backes, E. (eds) *Reducing Racial Inequality in Crime and Justice: Science, Practice,*

*and Policy* (National Academies Press, 2022); https://doi.org/10.17226/26705.

2. Carson, E. A. *Prisoners in 2019* No. 255115 (Bureau of Justice Statistics, 2020); https://bjs.ojp.gov/library/publications/prisoners-2019.

3. Carson, E. A. *Prisoners in 2020 – Statistical Tables* No. 302776 (Bureau of Justice Statistics, 2021); https://bjs.ojp.gov/library/publications/prisoners-2020-statistical-tables.

4. LeMasters, K. et al. Carceral epidemiology: mass incarceration and structural racism during the COVID-19 pandemic. *Lancet Public Health* **7**, e287–e290 (2022).

5. Gramlich, J. The gap between the number of blacks and whites in prison is shrinking. *Pew Research Center* https://www.pewresearch.org/fact-tank/2019/04/30/shrinking-gap-between-number-of-blacks-and-whites-in-prison/ (2019).

6. Klein, B. Dataset on incarcerated populations in the United States. Version 1.0.1. Zenodo https://doi.org/10.5281/zenodo.7675566 (2023).

7. Travis, J., Western, B. & Redburn S. (eds) *The Growth of Incarceration in the United States: Exploring Causes and Consequences* (National Research Council, National Academies Press, 2014); https://doi.org/10.17226/18613.

8. Davis, A. J. *Policing the Black Man: Arrest, Prosecution, and Imprisonment* (Pantheon Books, 2017).

9. Western, B. *Punishment and Inequality in America* (Russell Sage Foundation, 2006).

10. Hinton, E. & Cook, D. The mass criminalization of Black Americans: a historical overview. *Annu. Rev. Criminol.* **4**, 261–286 (2021).

11. Harrell, E. & Davis, E. *Contacts between Police and the Public, 2018 – Statistical Tables* No. 255730 (Bureau of Justice Statistics, 2020); https://bjs.ojp.gov/content/pub/pdf/cbpp18st.pdf.

12. Sawyer, W. How race impacts who is detained pretrial. *Prison Policy Initiative Briefing* https://www.prisonpolicy.org/blog/2019/10/09/pretrial_race/ (2019).

13. Subramanian, R., Digard, L., Washington, II M. & Sorage, S. *In the Shadows: A Review of the Research on Plea Bargaining* (Vera Institute of Justice, 2020); https://www.vera.org/downloads/publications/in-the-shadows-plea-bargaining.pdf.

14. *Demographic Differences in Sentencing: An Update to the 2012 Booker Report* (United States Sentencing Commission, 2018); https://www.ussc.gov/research/research-reports/demographic-differences-sentencing.

15. Edkins, V. A. & Redlich, A. D. *A System of Pleas: Social Sciences Contributions to the Real Legal System* (Oxford Univ. Press 2019).

16. Bonczar, T. & Beck, A. *Lifetime Likelihood of Going to State or Federal Prison* (Bureau of Justice Statistics, 1997); https://bjs.ojp.gov/library/publications/lifetime-likelihood-going-state-or-federal-prison.

17. *Criminal Justice Facts* (The Sentencing Project, 2021); https://www.sentencingproject.org/criminal-justice-facts/.

18. Duchin, M. & Spencer, D. Models, race, and the law. *Yale Law J.* **130**, 744–797 (2021).

19. Garland, B. E., Spohn, C. & Wodahl, E. J. Racial disproportionality in the American prison population: using the Blumstein method to address the critical race and justice issue of the 21st century. *Justice Policy J.* **5** (2) (Center on Juvenile and Criminal Justice, 2008); https://cdn2.rygn.io/cjcj/import/documents/racial_disproportionality.pdf.

20. Federal judges reinventing the jury trial during pandemic *United States Courts* https://www.uscourts.gov/news/2020/08/27/federal-judges-reinventing-jury-trial-during-pandemic (2020).

21. Baldwin, J. M., Eassey, J. M. & Brooke, E. J. Court operations during the COVID-19 pandemic. *Am. J. Crim. Justice* **45**, 743–758 (2020).

22. Daftary-Kapur, T., Henderson, K. S. & Zottoli, T. M. COVID-19 exacerbates existing system factors that disadvantage defendants: findings from a national survey of defense attorneys. *Law Hum. Behav.* **45**, 81–96 (2021).

23. Franco-Paredes, C. et al. Decarceration and community re-entry in the COVID-19 era. *Lancet Infect. Dis.* **21**, e11–e16 (2021).

24. *Legislature Approves Authority for Virginia DOC to Release Some Offenders Early During Pandemic* (Virginia Department of Corrections, 2020); https://vadoc.virginia.gov/media/1623/vadoc-financial-annual-mis-report-2020.pdf.

25. Kaminski, K. *MDOC Prison Population Update* (2020); https://www.house.mi.gov/hfa/PDF/Corrections/Corrections_Subcmte_Testimony_MDOC_Population_6-8-20.pdf.

26. *Response to COVID-19 – Early Release* (Utah Board of Pardons and Parole, 2020); https://bop.utah.gov/index.php/home-top-public-menu/2-uncategorised/165-bopp-coronavirus-information.

27. *Modified EPA Update* (Arkansas Board of Corrections, 2020); https://ssl-adc.ark.org/modified-epa-update.

28. Muller, C. & Roehrkasse, A. F. Racial and class inequality in U.S. incarceration in the early twenty-first century. *Social Forces* https://doi.org/10.1093/sf/soab141 (2021).

29. Sabol, W. J. & Johnson, T. L. *Justice System Disparities: Black-White National Imprisonment Trends, 2020-2022* (Council on Criminal Justice, 2022); https://counciloncj.foleon.com/reports/racial-disparities/.

30. *National Corrections Reporting Program, 1991-2019: Selected Variables* (United States Bureau of Justice Statistics, 2021); https://doi.org/10.3886/ICPSR38048.v1.

31. Cannon, R. T. Sick deal: injustice and plea bargaining during COVID-19. *J. Crim. Law Criminol.* https://doi.org/10.2139/ssrn.3632508 (2020).

32. Johnson, T. Crisis and coercive pleas. *J. Crim. Law Criminol.* https://doi.org/10.2139/ssrn.3632052 (2020).

33. Wildeman, C. & Wang, E. A. Mass incarceration, public health, and widening inequality in the USA. *Lancet* **389**, 1464–1474 (2017).

34. Beaudry, G. et al. Managing outbreaks of highly contagious diseases in prisons: a systematic review. *BMJ Global Health* https://doi.org/10.1136/bmjgh-2020-003201 (2020).

35. Macmadu, A. et al. COVID-19 and mass incarceration: a call for urgent action. *Lancet Public Health* **5**, e571–e572 (2020).

36. Hawks, L., Woolhandler, S. & McCormick, D. COVID-19 in prisons and jails in the United States. *JAMA Intern. Med.* **180**, 1041–1042 (2020).

37. Wang, E. A., Western, B. & Berwick, D. M. COVID-19, decarceration, and the role of clinicians, health systems, and payers: a report from the National Academy of Sciences, Engineering, and Medicine. *JAMA* **324**, 2257–2258 (2020).

38. Wang, E. A., Western, B., Backes, E. P. & Schuck, J. (eds) *Decarcerating Correctional Facilities during COVID-19: Advancing Health, Equity, and Safety* (National Academies of Sciences, Engineering, and Medicine, National Academies Press, 2020); https://doi.org/10.17226/25945.

39. Abraham, L. A., Brown, T. C. & Thomas, S. A. How COVID-19's disruption of the U.S. correctional system provides an opportunity for decarceration. *Am. J. Crim. Justice* **45**, 780–792 (2020).

40. Lacoste, A. M. B., Tyagi, E. & Johnson, H. *Fast, Frequent, and Widespread: COVID-19 Outbreaks inside Federal Prisons* (COVID

Behind Bars, 2021); https://uclacovidbehindbars.org/assets/federalprisonoutbreaks.pdf.

41. Nowotny, K. M., Bailey, Z. & Brinkley-Rubinstein, L. The contribution of prisons and jails to US racial disparities during COVID-19. *Am. J. Public Health* **111**, 197–199 (2021).

42. St. John, P. California to release 3,500 inmates early as coronavirus spreads inside prisons. *Los Angeles Times* (31 March 2020); https://www.latimes.com/california/story/2020-03-31/coronavirus-california-release-3500-inmates-prisons.

43. *Population Estimates, July 1, 2019, – Race and Hispanic Origin* (United States Census Bureau, 2019); https://www.census.gov/quickfacts/fact/table/US/PST045219.

44. Roberts, S., Peirce, J., Kajeepeta, S. & Bailey, M. *A Technical Guide to Jail Data Analysis* (Vera Institute of Justice, 2021); https://www.vera.org/downloads/publications/a-technical-guide-to-jail-data-analysis-report.pdf.

45. *COVID-19 Response Inmate Early Release Plan* (Virginia Department of Corrections, 2020); https://vadoc.virginia.gov/media/1506/vadoc-covid19-early-release-plan.pdf.

46. Skeem, J., Scurich, N. & Monahan, J. Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law Hum. Behav.* **44**, 51–59 (2020).

47. Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).

48. Lin, Z. J., Jung, J., Goel, S. & Skeem, J. The limits of human predictions of recidivism. *Sci. Adv.* **6**, eaaz0652 (2020).

49. Barabas, C., Dinakar, K. & Doyle, C. The problems with risk assessment tools. *New York Time*s (17 July 2019); https://www.nytimes.com/2019/07/17/opinion/pretrial-ai.html.

50. *National Landscape* (Mapping Pretrial Injustice, 2021); https://pretrialrisk.com/national-landscape/.

51. Xiang, A., Fogliato, R. & Chouldechova, A. Why PATTERN should not be used: the perils of using algorithmic risk assessment tools during COVID-19. *Partnership on AI* https://partnershiponai.org/why-pattern-should-not-be-used-the-perils-of-using-algorithmic-risk-assessment-tools-during-covid-19/ (2020).

52. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine bias. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).

53. Althouse, B. M. et al. The unintended consequences of inconsistent pandemic control policies. Preprint at *medRxiv* https://doi.org/10.1101/2020.08.21.20179473 (2020).

54. Flanagin, A., Frey, T. & Christiansen, S. L. & AMA Manual of Style Committee. Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA* **326**, 621–627 (2021).

55. Boehmer, U. et al. Self-reported vs administrative race/ethnicity data and study results. *Am. J. Public Health* **92**, 1471–1472 (2002).

# Acknowledgements

We thank H. Hartle, S. McCabe, T. LaRock, R. Gallagher, M. Frank, R. Rohlfs, E. Ross, B. Terry, S. Goel, D. Allen and members of the Northeastern University NULab for helpful conversations and tips with constructing the dataset. We thank members of the Justice Collaboratory at the Yale Law School, the Institute on Policing, Incarceration & Public Safety at The Hutchins Center for African & African American Research, Harvard University, the MLK Visiting Scholars Program at the Massachusetts Institute of Technology and the Vera Institute for Justice for support, helpful exchanges on the subject matter and feedback on the manuscript. Finally, we thank the Santa Fe Institute, New York University, University of Pittsburgh and the Yale School of Public Health for invitations to seminars and colloquia, at which aspects of this study were discussed.

# Author information

## Authors and Affiliations

1. Network Science Institute, Northeastern University, Boston, MA, USA

   Brennan Klein, Nitish Kaza, Arush Sharma, Tina Eliassi-Rad & Samuel V. Scarpino

2. Institute on Policing, Incarceration & Public Safety, The Hutchins Center for African & African American Research, Harvard University, Cambridge, MA, USA

   Brennan Klein & Elizabeth Hinton

3. Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

   C. Brandon Ogbunugafor

4. Public Health Modeling Unit, Yale School of Public Health, New Haven, CT, USA

   C. Brandon Ogbunugafor

5. Santa Fe Institute, Santa Fe, NM, USA

   C. Brandon Ogbunugafor, Tina Eliassi-Rad & Samuel V. Scarpino

6. Vermont Complex Systems Center, University of Vermont, Burlington, VT, USA

   C. Brandon Ogbunugafor, Tina Eliassi-Rad & Samuel V. Scarpino

7. Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA

   C. Brandon Ogbunugafor

8. Department of History, Yale University, New Haven, CT, USA

Benjamin J. Schafer & Elizabeth Hinton

9. College of Engineering, Northeastern University, Boston, MA, USA

Zarana Bhadricha & Preeti Kori

10. Roux Institute, Northeastern University, Boston, MA, USA

Jim Sheldon & Samuel V. Scarpino

11. SEICHE Center for Health and Justice, Yale School of Medicine, New Haven, CT, USA

Emily A. Wang

12. Department of Medicine, Yale School of Medicine, New Haven, CT, USA

Emily A. Wang

13. Justice Collaboratory, Yale Law School, New Haven, CT, USA

Emily A. Wang & Elizabeth Hinton

14. Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA

Tina Eliassi-Rad & Samuel V. Scarpino

15. The Institute for Experiential AI, Northeastern University, Boston, MA, USA

Tina Eliassi-Rad & Samuel V. Scarpino

16. Department of Health Sciences, Northeastern University, Boston, MA, USA

Samuel V. Scarpino

17. Department of African American Studies, Yale University, New Haven, CT, USA

Elizabeth Hinton

18. Yale Law School, New Haven, CT, USA

Elizabeth Hinton

### Contributions

B.K., C.B.O., S.V.S. and E.H. conceived the project. C.B.O., S.V.S. and E.H. directed the project. B.K. directed the construction of the data science pipeline. B.K., B.J.S., Z.B., P.K., J.S., A.S. and N.K. collected data. B.K., C.B.O., B.J.S. and S.V.S. conducted analyses. B.K., C.B.O., E.A.W., T.E.-R., S.V.S. and E.H. interpreted and integrated the results. B.K., C.B.O., B.J.S., E.A.W., T.E.-R., S.V.S. and E.H. contributed to researching, writing and editing the final manuscript.

### Corresponding authors

Correspondence to Brennan Klein, C. Brandon Ogbunugafor, Samuel V. Scarpino or Elizabeth Hinton.

## Ethics declarations

### Competing interests

The authors declare no competing interests.

## Peer review

### Peer review information

*Nature* thanks Joshua Page, John Pfaff, Vesla Weaver and the other, anonymous, reviewer(s) for their contribution to the peer review of this

work. Peer reviewer reports are available.

## Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Supplementary information

### Supplementary Information

This file contains Supplementary Sections 1–5 and includes Supplementary Figs. 1–23 and Supplementary Tables 1–7.

### Reporting Summary

### Peer Review File

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

# About this article

## Cite this article

Klein, B., Ogbunugafor, C.B., Schafer, B.J. *et al.* COVID-19 amplified racial disparities in the US criminal legal system. *Nature* **617**, 344–350 (2023). https://doi.org/10.1038/s41586-023-05980-2

- Received: 11 January 2022
- Accepted: 20 March 2023
- Published: 19 April 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05980-2

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

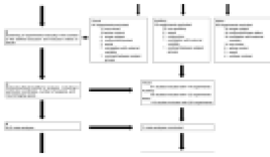
Provided by the Springer Nature SharedIt content-sharing initiative

## COVID pandemic increased racial disparities in US prison populations

- Jessica M. Eaglin

Nature News & Views 19 Apr 2023

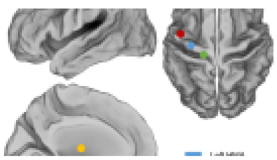
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05980-2



## Delineating visual, auditory and motor regions in the human brain with functional neuroimaging: a BrainMap-based meta-analytic synthesis

11 May 2021

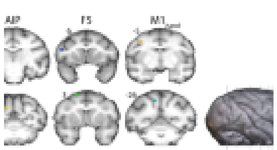
Marisa K. Heckner, Edna C. Cieslik, … Robert Langner



## Hand, foot and lip representations in primary sensorimotor cortex: a high-density electroencephalography study

19 December 2019

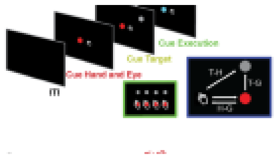
Mingqi Zhao, Marco Marino, … Dante Mantini




## Shared functional connectivity between the dorso-medial and dorso-ventral streams in macaques

29 October 2020

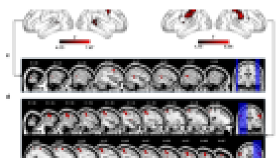
R. Stefan Greulich, Ramina Adam, … Hansjörg Scherberger



## [The human primary somatosensory cortex encodes imagined movement in the absence of sensory information]

11 December 2020

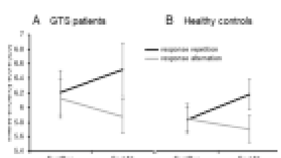
Matiar Jafari, Tyson Aflalo, … Richard Alan Andersen



## [Domain-specific functional coupling between dorsal and ventral systems during action perception]

03 December 2020

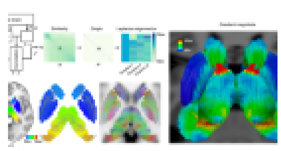
Huichao Yang, Chenxi He, … Yanchao Bi



## [Somatosensory perception–action binding in Tourette syndrome]

28 June 2021

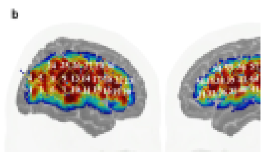
Julia Friedrich, Henriette Spaleck, … Alexander Münchau



[Topographic organization of the human subcortex unveiled with functional connectivity gradients]

28 September 2020

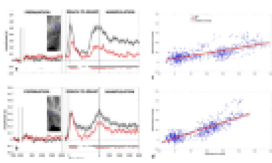
Ye Tian, Daniel S. Margulies, … Andrew Zalesky



[Brain correlates of motor complexity during observed and executed actions]

03 July 2020

Xinge Li, Manon A. Krol, … Meryem A. Yücel



[Action execution and action observation elicit mirror responses with the same temporal profile in human SII]

20 February 2020

Maria Del Vecchio, Fausto Caruana, … Pietro Avanzini

- Article
- Open Access
- [Published: 19 April 2023]

# A somato-cognitive action network alternates with effector regions in motor cortex


- Evan M. Gordon ,
- Roselyne J. Chauvin,
- Andrew N. Van,
- Aishwarya Rajesh,
- Ashley Nielsen,
- Dillan J. Newbold,
- Charles J. Lynch,
- Nicole A. Seider,
- Samuel R. Krimmel,
- Kristen M. Scheidter,
- Julia Monk,
- Ryland L. Miller,
- Athanasia Metoki,
- David F. Montez,
- Annie Zheng,
- Immanuel Elbau,
- Thomas Madison,
- Tomoyuki Nishino,
- Michael J. Myers,
- Sydney Kaplan,
- Carolina Badke D’Andrea,
- Damion V. Demeter,
- Matthew Feigelis,
- Julian S. B. Ramirez,
- Ting Xu,
- Deanna M. Barch,
- Christopher D. Smyser,
- Cynthia E. Rogers,
- Jan Zimmermann,
- Kelly N. Botteron,

- John R. Pruett,
- Jon T. Willie,
- Peter Brunner,
- Joshua S. Shimony,
- Benjamin P. Kay,
- Scott Marek,
- Scott A. Norris,
- Caterina Gratton,
- Chad M. Sylvester,
- Jonathan D. Power,
- Conor Liston,
- Deanna J. Greene,
- Jarod L. Roland,
- Steven E. Petersen,
- Marcus E. Raichle,
- Timothy O. Laumann,
- Damien A. Fair &
- …
- Nico U. F. Dosenbach

Show authors

*Nature* **volume 617**, pages 351–359 (2023)

- 102k Accesses
- 1 Citations
- 1691 Altmetric
- Metrics details

# Abstract

Motor cortex (M1) has been thought to form a continuous somatotopic homunculus extending down the precentral gyrus from foot to face representations[1,2], despite evidence for concentric functional zones[3] and

maps of complex actions[4]. Here, using precision functional magnetic resonance imaging (fMRI) methods, we find that the classic homunculus is interrupted by regions with distinct connectivity, structure and function, alternating with effector-specific (foot, hand and mouth) areas. These inter-effector regions exhibit decreased cortical thickness and strong functional connectivity to each other, as well as to the cingulo-opercular network (CON), critical for action[5] and physiological control[6], arousal[7], errors[8] and pain[9]. This interdigitation of action control-linked and motor effector regions was verified in the three largest fMRI datasets. Macaque and pediatric (newborn, infant and child) precision fMRI suggested cross-species homologues and developmental precursors of the inter-effector system. A battery of motor and action fMRI tasks documented concentric effector somatotopies, separated by the CON-linked inter-effector regions. The inter-effectors lacked movement specificity and co-activated during action planning (coordination of hands and feet) and axial body movement (such as of the abdomen or eyebrows). These results, together with previous studies demonstrating stimulation-evoked complex actions[4] and connectivity to internal organs[10] such as the adrenal medulla, suggest that M1 is punctuated by a system for whole-body action planning, the somato-cognitive action network (SCAN). In M1, two parallel systems intertwine, forming an integrate–isolate pattern: effector-specific regions (foot, hand and mouth) for isolating fine motor control and the SCAN for integrating goals, physiology and body movement.

# Main

Beginning in the 1930s, Penfield and colleagues mapped human M1 with direct cortical stimulation, eliciting movements from about half of sites, mostly of the foot, hand and mouth[1]. Although representations for specific body parts overlapped substantially[11], these maps gave rise to the textbook view of M1 organization as a continuous homunculus, from head to toe.

In non-human primates, organizational features inconsistent with the motor homunculus have been described. Structural connectivity studies divided M1 into anterior, gross motor, ‘old’ M1 (few direct projections to spinal motor neurons) and posterior, fine motor, ‘new’ M1[12,13] (many direct

motoneuronal projections). Non-human primate stimulation studies showed the body to be represented in anterior M1[14], and the motor effectors (tail, foot, hand and mouth) in posterior M1. Such studies also suggested that the limbs are represented in concentric functional zones progressing from the digits at the centre, to the shoulders on the periphery[3]. Moreover, stimulations could elicit increasingly complex and multi-effector actions when moving from posterior to anterior M1[4].

During natural behaviour, voluntary movements are part of goal-directed actions, initiated and controlled by executive regions in the CON[5]. Neural activity preceding voluntary movements can first be detected in the rostral cingulate zone[15] within dorsal anterior cingulate cortex (dACC), then in the pre-supplementary motor area (pre-SMA) and supplementary motor area[16] (SMA), followed by M1. These regions all project to the spinal cord[17], with M1 as the main transmitter of motor commands down the corticospinal tract[18]. Efferent motor copies are received by primary somatosensory cortex[19] (S1), cerebellum[20] and striatum[21] for online correction, learning[20] and inhibition of competing movements[22]. Tracer injections in non-human primates demonstrated projections from anterior M1/CON to internal organs (such as adrenal medulla) for preparatory sympathetic arousal in anticipation of action[10]. Post-movement error and pain signals are relayed primarily to insular and cingulate regions of the CON, which update future action plans[8,9].

Resting-state functional connectivity (RSFC) fMRI noninvasively maps the brain’s functional networks[23]. Precision functional mapping (PFM) studies rely on large amounts of multi-modal data (such as RSFC and tasks) to map individual-specific brain organization in the greatest possible detail[24,25]. Early PFM studies identified separate foot, hand and mouth M1 regions[24] with their respective cerebellar and striatal targets[26,27]. These foot, hand and mouth motor circuits were characterized by strong within-circuit connectivity and effector specificity in task fMRI[24], consistent with myeloarchitectonic evidence for distinct cortical fields[28]. However, these circuits were relatively isolated and did not include functional connections with control networks such as CON that could support the integration of movement with global behavioural goals. A recent study showed that

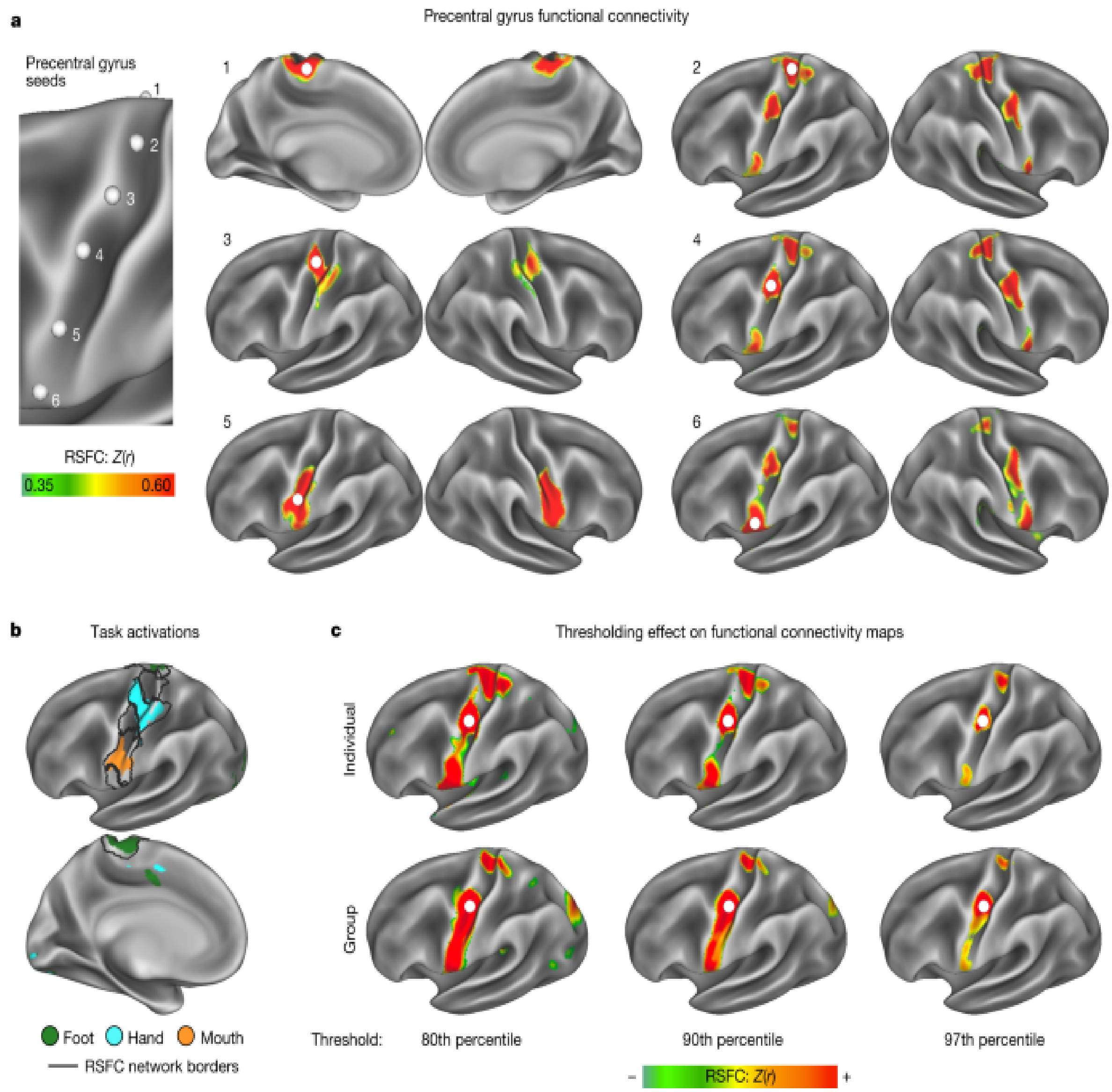
prolonged dominant arm immobilization strengthened functional connectivity between disused M1 and the CON[29,30], suggesting that the role of CON may extend beyond abstract action control and into movement coordination.

Here we used the latest iteration of PFM with higher resolution (2.4 mm) and greater amounts of fMRI (RSFC: 172–1,813 min per participant; task: 353 min per participant), and diffusion data, to map M1 and its connections with the highest detail. The results were verified in group-averaged data from the three largest fMRI studies (Human Connectome Project (HCP), Adolescent Brain Cognitive Development (ABCD) study, UK Biobank (UKB); total *n* of approximately 50,000). Furthermore, we placed our findings in cross-species (macaque versus human), developmental (neonate, infant, child and adult) and clinical (perinatal stroke) contexts using PFM data.

# Two networks alternate in motor cortex

Advanced PFM revealed connectivity that differed markedly from the canonical homuncular organization of M1. Two contrasting patterns of functional connectivity alternated in M1 (Fig. 1a and Supplementary Video 1). The expected pattern, as previously described for M1 foot, hand and mouth representations[24,31], comprised three regions (per hemisphere) for which cortical connectivity was restricted to homotopic contralateral M1 and adjacent S1 (Fig. 1a, seeds 1, 3 and 5). This set of RSFC-defined regions corresponded with task-evoked activity during foot, hand and tongue movements (Fig. 1b; see Extended Data Fig. 1d for other participants).

**Fig. 1: Precision functional mapping of primary motor cortex.**




**a**, RSFC seeded from a continuous line of cortical locations in the left precentral gyrus in a single exemplar participant (P1; 356 min resting-state fMRI). The six exemplar seeds shown represent all distinct connectivity patterns observed (see Supplementary Video 1 for complete mapping). Functional connectivity seeded from these locations illustrated classical M1 connectivity of regions representing the foot (1), hand (3) and mouth (5), as well as an interdigitated set of strongly interconnected regions (2, 4 and 6). See Extended Data Fig. 1a and Supplementary Video 2 for all highly sampled participants, Extended Data Fig. 1b for within-participant replications, and Extended Data Fig. 1c for group-averaged data. **b**, Discrete functional networks were demarcated using a whole-brain, data-driven,

hierarchical approach (Methods) applied to the resting-state fMRI data, which defined the spatial extent of the networks observed in Fig. 1 (black outlines). Regions defined by RSFC were functionally labelled using a classic block-design fMRI motor task involving separate movement of the foot, hand and tongue (following ref. [31]; see ref. [29] for details). The map illustrates the top 1% of vertices activated by movement of the foot, hand and mouth in the exemplar participant (P1; see Extended Data Fig. 1d for other participants). **c**, The inter-effector connectivity pattern became more distinct from surrounding effector-specific motor regions as connectivity thresholding increased from the 80th to the 97th percentile. RSFC thresholds required to detect the inter-effector pattern were lower in individual-specific data (top) than in group-averaged data (bottom; ABCD study, $n = 3{,}928$).

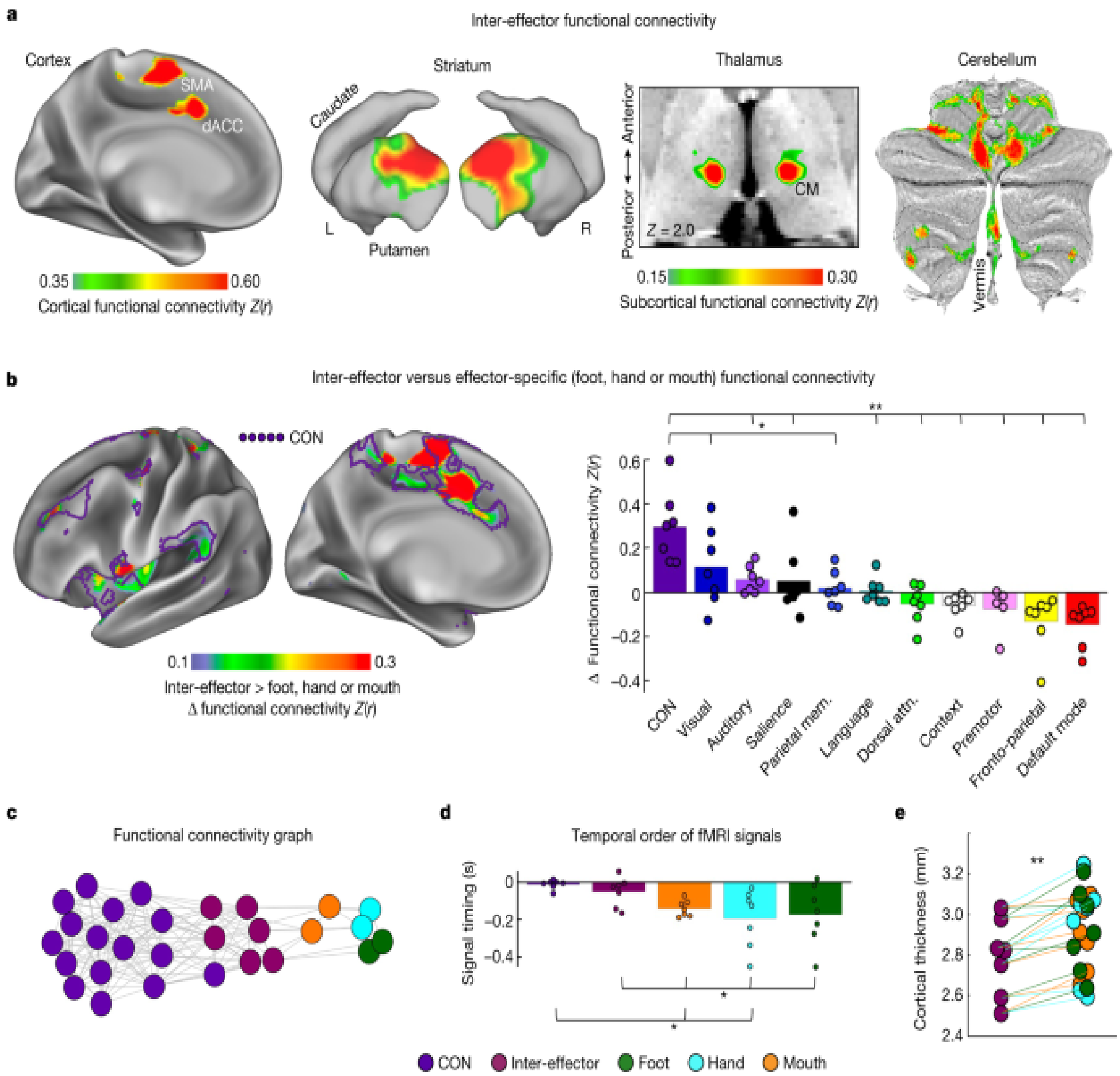
Interleaved between the known foot, hand and mouth M1 regions lay three areas that were strongly functionally connected to each other, both contralaterally and ipsilaterally, forming a previously unrecognized interdigitated chain down the precentral gyrus (Fig. 1a, seeds 2, 4, 6). The motif of three M1 inter-effector regions was observed in every highly sampled adult (Extended Data Fig. 1a and Supplementary Table 1) and replicated within-individual in separate data from the same participants (Extended Data Fig. 1b). Of note, the inter-effector pattern was also evident in all large *n* group-averaged data (UKB ($n = 4{,}000$), ABCD ($n = 3{,}928$), HCP ($n = 812$) and WU120 ($n = 120$); Extended Data Fig. 1c). The M1 inter-effector functional connectivity motif was most apparent in individual-specific maps, but once recognized, was also clearly identifiable in group-averaged data when visualized using stringent connectivity thresholds (Fig. 1c).

The inter-effector regions were evident relatively early in development. Whereas PFM data from a human newborn did not reveal the inter-effector motif, it was detectable in an 11-month-old infant, and was almost adult-like in a 9-year-old child (Extended Data Fig. 2a–e). Inter-effector regions could even be identified in an individual with preserved motor function despite suffering severe bilateral perinatal strokes that destroyed large portions of M1 (Extended Data Fig. 2f; see ref. [32] for clinical details).

## Inter-effectors link to control network

In addition to being interconnected, the three inter-effector regions were functionally connected to multiple regions of the CON, thought to be important for goal-oriented cognitive control. Connectivity was very strong with SMA and a region in dACC[15] (caudal cingulate zone) (Fig. 2a; see Extended Data Fig. 3 and Supplementary Table 1 for all participants) but was also evident with anterior prefrontal cortex (aPFC) and insula (Supplementary Fig. 1). In striatum, inter-effector regions were most strongly connected to dorsolateral putamen. In thalamus, connectivity peaked in the centromedian (CM) nucleus, with additional strong connectivity observed in ventral intermediate (VIM), ventral posteriomedial (VPM) and ventral posterior inferior (VPI) nuclei. Inter-effector regions were strongly connected to cerebellar areas (Fig. 2a) surrounding but distinct from effector-specific cerebellar regions (Extended Data Fig. 1e).

**Fig. 2: Functional connectivity and cortical thickness of the M1 inter-effector motif.**




**a**, Brain regions with the strongest functional connectivity to the left middle inter-effector region (exemplar seed) in cortex, striatum, thalamus (horizontal slice; CM nucleus) and cerebellum (flat map) in the exemplar participant (P1). See Extended Data Fig. 3 for other participants. **b**, Left, brain regions more strongly functionally connected to inter-effectors than to any foot, hand or mouth regions (P1; Supplementary Fig. 2a for other participants). Purple outlines show the CON (individual-specific). Central sulcus is masked as it exhibits large differences by definition. Right, connectivity was calculated between every network and both the inter-effector and effector-specific M1 regions. The plot shows the smallest difference between inter-effector and any effector-specific connectivity, averaged across participants. This difference was larger for CON than for

any other network (two-tailed paired $t$-tests, *$P < 0.05$, FDR-corrected; **$P < 0.01$, FDR-corrected). Coloured circles represent individual participants. **c**, Inter-network relationships visualized in network space using a spring-embedding plot, in which connected regions are pulled together and disconnected regions are pushed apart. Connecting lines indicate a functional connection ($Z(r) > 0.2$) (P1; see Supplementary Fig. 2b for all participants). **d**, Inter-effector and effector-specific regions were tested for systematic differences in the temporal ordering of their infra-slow fMRI signals[34] (<0.1 Hz). The plot shows signal ordering in CON, inter-effector and effector-specific regions, averaged across participants (standard error bars; two-tailed paired $t$-test *$P < 0.05$, uncorrected). Coloured circles represent individual participants. Prior electrophysiology work suggests that later infra-slow activity (here, CON) corresponds to earlier delta-band (0.5–4 Hz) activity[35]. **e**, In each participant (filled circles), inter-effector regions exhibited lower cortical thickness than all effector-specific regions (two-tailed paired $t$-test **$P \leq 0.01$, FDR-corrected). Attn., attention; mem., memory.

In all highly sampled individuals ($n = 7$), the inter-effector regions had stronger connections to CON than did any of the foot, hand or mouth regions (Fig. 2b; Supplementary Fig. 2a for all participants); across participants: all two-tailed paired $t > 4.75$, $P < 0.01$ false discovery rate (FDR) corrected, for inter-effector versus foot, versus hand, and versus mouth (Extended Data Fig. 4a). The inter-effector versus foot, hand and mouth difference was larger for CON than for any of the other 10 networks (all two-tailed paired $t > 3.5$; all $P < 0.05$, FDR-corrected; Fig. 2b). In network space, inter-effector regions were positioned between CON and the foot, hand and mouth regions (Fig. 2c; Supplementary Fig. 2b for all participants). Inter-effector regions were also more strongly connected to: middle insula, known to process pain[9] and interoceptive signals[33] (Extended Data Fig. 4b; all two-tailed paired $t > 2.7$; all $P < 0.05$, FDR-corrected); lateral cerebellar lobule V and vermis Crus II, lobule VIIb and lobule VIIIa (all two-tailed paired $t > 3.7$, all $P < 0.05$, FDR-corrected; Extended Data Fig. 4c); dorsolateral putamen, critical for motor function (all two-tailed paired $t > 3.7$; all $P < 0.01$, FDR-corrected, Extended Data Fig. 4d); and sensory-motor regions of thalamus (VIM, CM and VPM; all two-tailed paired $t > 3.0$, all $P < 0.03$, FDR-corrected; Extended Data Fig. 4e–g). Searching for differences between the three inter-

effector regions revealed that the middle inter-effector region consistently exhibited stronger functional connectivity to extrastriate visual cortex than did the superior and inferior inter-effector regions (Extended Data Fig. 5; Supplementary Fig. 3 for all participants).

Comparing the relative timing of resting-state fMRI signals (lag structure[34]) showed that infra-slow (<0.1 Hz) fMRI signals in both the CON and the inter-effector network lagged behind those in effector-specific regions (Fig. 2d; CON versus foot: two-tailed paired $t = 2.38$, $P = 0.055$, uncorrected; versus hand and mouth: all two-tailed paired $t > 2.84$, all $P < 0.03$, uncorrected; inter-effector versus foot, hand and mouth: all two-tailed paired $t > 2.5$, all $P < 0.05$, uncorrected). Inter-regional lags in infra-slow (<0.1 Hz) signals are associated with propagation of higher-frequency delta activity (0.5–4 Hz) in the opposite direction[35], suggesting that high-frequency signals may occur earlier in CON than in M1—consistent with electrical recordings during voluntary movement[36]—but that such signals reach the inter-effectors earlier than the foot, hand and mouth regions.

As expected, the M1 foot, hand and mouth regions were strongly functionally connected with adjacent S1 (Fig. 1a and Extended Data Fig. 6a), consistent with known functional connections between M1 and S1[37]. By contrast, inter-effector regions exhibited lower connectivity with adjacent S1 (Extended Data Fig. 4h; all two-tailed paired $t > 3.2$, all $P < 0.02$, FDR-corrected). More specifically, inter-effector functional connectivity extended into the fundus of the central sulcus (Extended Data Fig. 6b; Brodmann area (BA) 3a), which represents proprioception[38], but not to the postcentral gyrus (BA1, BA2 and BA3b), representing cutaneous tactile stimuli.

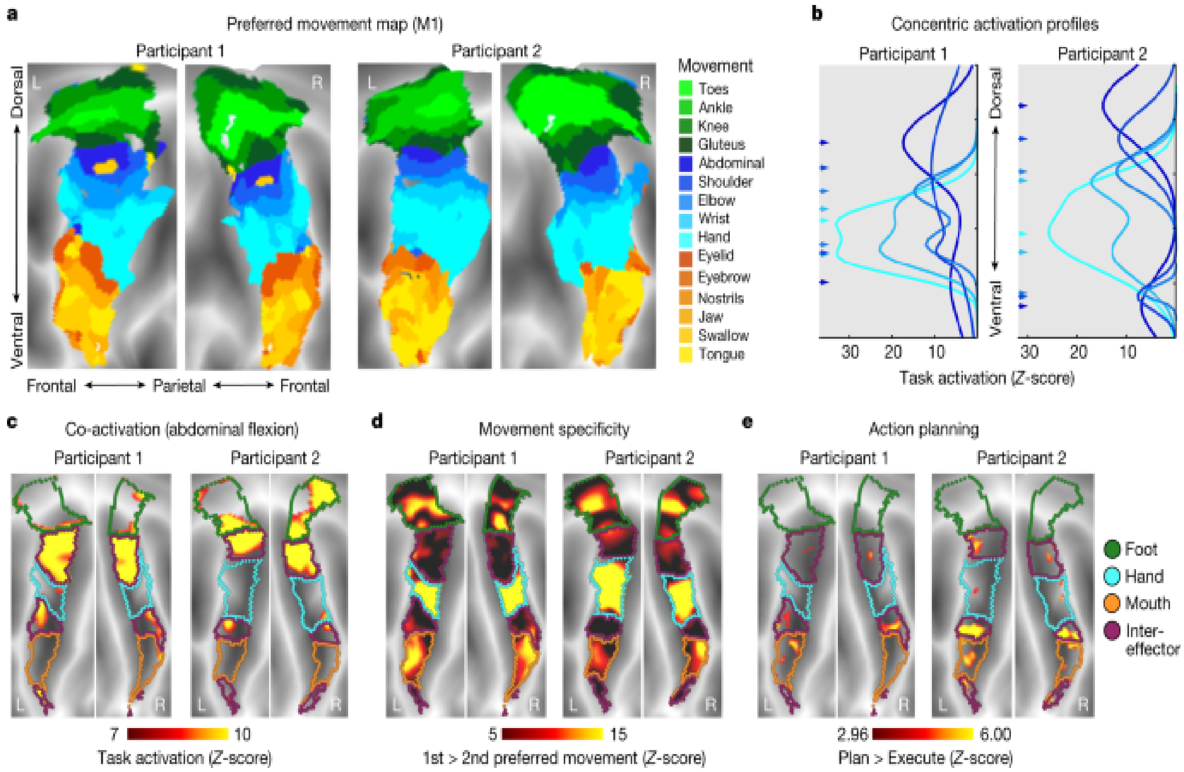
Convergent with these functional differences, metrics of brain structure systematically differed between inter-effector and effector-specific regions. Inter-effector regions exhibited lower cortical thickness (all two-tailed paired $t > 3.6$; all $P \leq 0.01$, FDR-corrected; Fig. 2e), more similar to prefrontal cortex[39], but higher fractional anisotropy (2 mm beneath cortex; all two-tailed paired $t > 5.3$; all $P < 0.05$, FDR-corrected; Extended Data Fig. 4j). Intracortical myelin content was higher in inter-effector regions than in foot regions (two-tailed paired $t = 6.8$, $P < 0.005$, FDR-corrected) but lower than in hand regions (two-tailed paired $t = 4.8$, $P < 0.005$, FDR-corrected;

Extended Data Fig. 4k), suggesting myeloarchitectonic differences similar to those described in ref. [28].

## Concentric motor and body-action zones

To better understand the functions of the inter-effector motif, we collected fMRI data during blocked performance of 25 different movements in 2 highly sampled individuals (64 runs; 244 min per participant) and during a novel event-related task with separate planning and execution phases for coordinated hand and foot movements (12 runs; 132 min per participant). According to the homuncular model of M1, activation when moving a given body part should exhibit a single peak within the precentral gyrus. If M1 is instead organized into concentric functional zones, all movements except those at the centres (that is, toes, fingers, tongue) should exhibit two peaks (above and below). Within each of the three effector-specific regions, the topography of preferred movements—the movement eliciting greatest activation in each vertex (Fig. 3a)—was more consistent with a concentric organization (distal-proximal; for example, toes in the centre, with surrounding concentric zones of ankle–knee–hip)[3] than with the canonical, linear toes-to-face homuncular model[1].

**Fig. 3: Individual-specific task activations in M1.**




**a**, Task fMRI activations (P1 and P2) during a movement task battery, including movement of the toes, ankles, knees, gluteus, abdominals, shoulders, elbows, hands, eyebrows, eyelids, tongue and swallowing (244 min per participant). Each cortical vertex is coloured according to the movement that elicited the strongest task activation (winner takes all) and is shown on a flattened representation of the cortical surface. Background shading indicates sulcal depths. **b**, Activation strength for each movement was computed along the dorsal–ventral axis within M1. A two-peak Gaussian curve was fitted to each movement activation (Methods). Fitted curves are shown for movement of abdominals, shoulder, elbow, wrist and hand. Peak locations (arrows on left) were arranged concentrically around the hand peak. See Extended Data Fig. 7 and Supplementary Fig. 4 for all movements. **c**, Inter-effector regions were co-activated during abdominal contraction. **d**, Inter-effector regions exhibited more generalized evoked activity during movements. Movement specificity was computed as the activation difference between the first- and second-most preferred movements for the six conditions that most activated each discrete region (toes, abdominal, hand, eyelid, tongue and swallowing). **e**, Event-related task fMRI data during an action planning task with separate planning and execution phases for movements of the hands and feet (Methods). M1

activity in the planning phase was higher than in the execution phase in the inter-effector but not the effector-specific regions.

To formally test for a concentric organization, we fit one- and two-peak Gaussian curves to the task activation profiles along the dorsomedial-to-ventrolateral axis of M1. Two-peak curve fits were significantly better for all movements ($F$-test for comparing models (Methods): all $F > 6.9$, all $P <$ 0.001, FDR-corrected) except hand in P2 ($F \cong 0$, $P \cong 1$) (Supplementary Fig. 4). The curve fits revealed concentric activation zones centred around activation peaks for distal movements (hand (Fig. 3b), toes and tongue (Supplementary Fig. 4)) and expanding outward to more proximal movements (shoulder, gluteus and jaw). Concentric rings of activation from separate foot, hand and mouth centres intersected in the superior and middle inter-effector regions (Extended Data Fig. 7).

Some movements requiring less fine motor control, such as isometric contraction of the abdominals (Fig. 3c) or raising the eyebrow co-activated multiple inter-effector regions and the CON (Extended Data Figs. 7 and 8a,b,e). By contrast, movements of the foot and hand only activated the corresponding effector-specific regions (Extended Data Figs. 7 and 8c–e). Unlike effector-specific regions, the inter-effectors exhibited weak movement specificity, with minimal activation differences between their preferred and non-preferred movements (Fig. 3d) and at least some activation observed across most movements (Extended Data Fig. 7).

To verify that inter-effector function is not specific to vocalization[40], we also collected task fMRI data while participants repeatedly made an ‘ee’ sound, to isolate movement of the larynx while minimizing respirations and jaw and tongue motion. We observed a dual laryngeal representation that was confined to the mouth area rather than extending into the inter-effector regions (Supplementary Fig. 5), consistent with ref. [41] and a concentric functional zone organization.

Regions in CON instantiate action plans, suggesting that the CON-to-inter-effector connection could carry general action planning signals. Across foot and hand movements in a novel coordination task, the inter-effectors showed greater activity during action planning than movement execution but the effector-specific regions did not (Fig. 3e), suggesting that the

implementation of action plans may be enabled in part by the inter-effector regions in M1.

## Macaque homologue of body/action network

To link these neuroimaging findings to decades of detailed motor mapping in non-human primates, we searched for inter-effector homologues in macaques using fMRI. Seeds placed in macaque M1 revealed foot, hand and mouth effector-specific functional connectivity patterns consistent with those seen in humans[24] (Extended Data Fig. 9, rows 2–4). Seeding putative CON homologues in dACC (see Supplementary Table 2 for seed locations), revealed strong connectivity with lateral frontal cortex, insula and supramarginal gyrus, similar to the human CON, and with two regions in anterior central sulcus potentially homologous to the superior and middle inter-effectors (Extended Data Fig. 9, row 1).

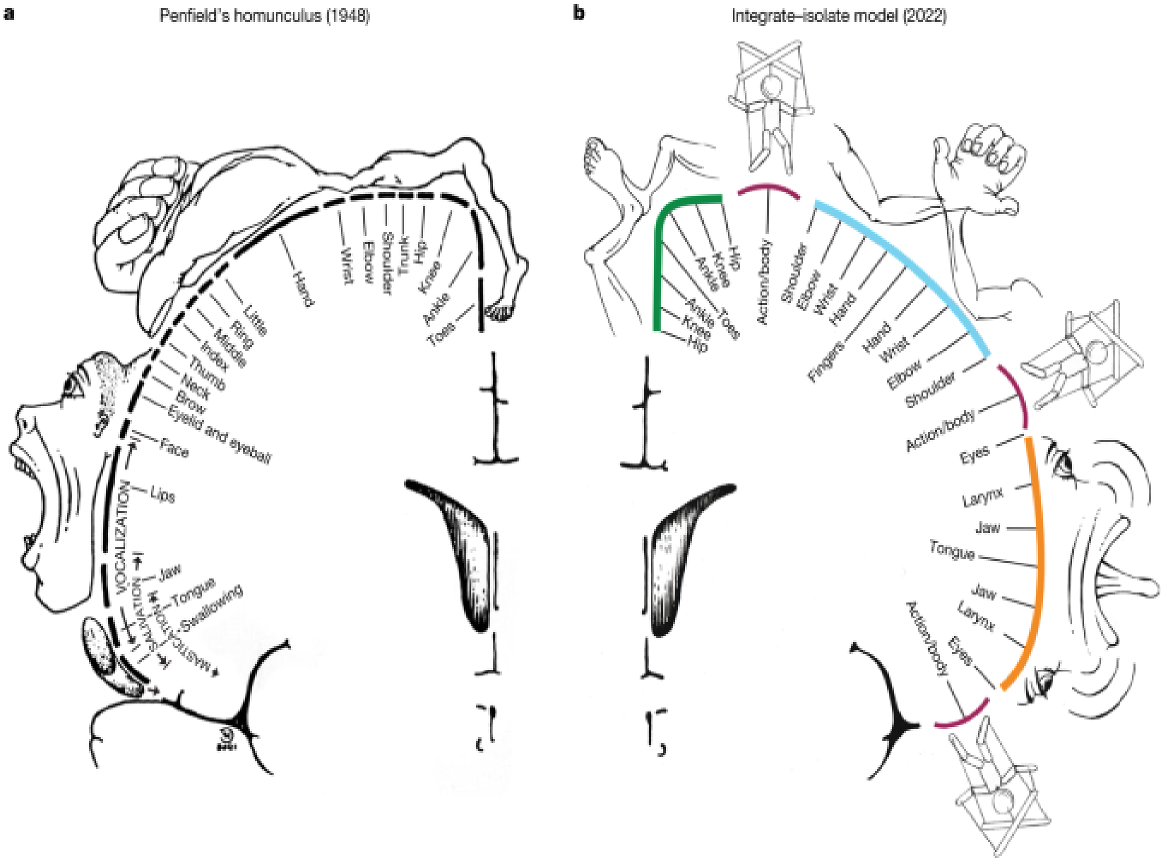
Distinct patterns of corticospinal connectivity are known to distinguish separable regions of macaque M1[12,13]. Phylogenetically newer, posterior M1 represents the effectors[14], projects contralaterally—mainly to the cervical and lumbar enlargements of the spinal cord[13]—and contains more projections that synapse directly onto muscle-innervating spinal neurons[12] for fine motor control. By contrast, older anterior M1 represents the body[14], projects bilaterally throughout the spinal cord[13], and connects to internal organs such as the adrenal medulla[10] and stomach[42]. Notably, the spatial distribution of adrenal connectivity[10] converges with the proposed inter-effector homologues and connected medial wall regions (SMA and dACC).

Direct stimulation studies in macaques have evoked complex, multi-effector actions by applying longer stimulation trains (500 ms) to motor cortex[4]. These actions range from feeding behaviours to climbing and defensive postures—movements that are purposeful and coordinated rather than isolated, involving integration of muscles across the classic foot, hand and mouth divisions. The inter-effector regions, which are connected to action planning areas (Fig. 2) and are active during a wide range of foot, hand and mouth movements (Extended Data Fig. 7), represent candidate human homologues to the macaque multi-effector action sites.

# Effector isolation versus action integration

Penfield conceptualized his direct stimulation findings in M1 as a continuous map of the human body—the homunculus—an organizational principle that has been dominant for almost 100 years (Fig. 4a). On the basis of novel and extant data, we instead propose a dual-systems, integrate–isolate model of behavioural control, in which effector isolating and whole-organism action implementation regions alternate (Fig. 4b). This model better fits the human imaging data presented here demonstrating contrasting structural, functional and connectivity patterns within M1 (Extended Data Fig. 4). The inter-effector patterning emerges in infancy and is preserved even in the presence of substantial perinatal cortical injury (Extended Data Fig. 2). In the integrate–isolate model, the regions for foot, hand and mouth fine motor skill are organized somatotopically as three concentric functional zones with distal parts of the effector (toes, fingers and tongue) at the centre and proximal ones (knee, shoulder and larynx) on the perimeter (Fig. 3a,b, Extended Data Fig. 7 and Supplementary Fig. 4). It has been suggested that this concentric organization extends to the ordering of fingers within the hand representation[43]. Effector-specific regions activate strongly for preferred movements and are commonly deactivated for non-preferred movements (Fig. 3d and Extended Data Fig. 7). The inter-effector regions at the edges of the effector zones coordinate with each other and with the CON (Extended Data Figs. 7 and 8; see also ref. [44]) to accomplish holistic, whole-body functions in the service of performing actions (Fig. 3e). The present work suggests that these functions include action implementation, as well as postural and gross motor control of axial muscles, and prior work in humans and non-human primates suggests that these circuits may also regulate arousal[7], coordinate breathing with speech and other complex actions[45], and control internal processes and organs (such as, blood pressure[6], stomach[42] and adrenal medulla[10]), consistent with circuits for whole-body, metabolic and physiological control. Minor connectivity (Extended Data Fig. 5) and activation (Extended Data Fig. 7) differences between the superior, middle and inferior inter-effector regions probably reflect some degree of functional specialization within this integrated system. The middle region's relatively stronger connectivity to visual cortex, for example, could suggest a potential role in hand–eye coordination during reach-and-grasp motions[4].

**Fig. 4: The interrupted homunculus, an integrate–isolate model of action and motor control.**




**a**, Penfield's classical homunculus (adapted from ref. [2]), depicting a continuous map of the body in primary motor cortex. **b**, In the integrate–isolate model of M1 organization, effector-specific—foot (green), hand (cyan) and mouth (orange)—functional zones are represented by concentric rings with proximal body parts surrounding the relatively more isolatable distal ones (toes, fingers and tongue). Inter-effector regions (maroon) sit at the intersecting points of these fields, forming part of a somato-cognitive action network for integrative, allostatic whole-body control. As with Penfield's original drawing, this diagram is intended to illustrate organizational principles, and must not be over-interpreted as a precise map.

Thus, the inter-effector system fulfils the role of a somato-cognitive action system (SCAN). The SCAN forms part of an integrated action control system, in conjunction with the CON's upstream executive control operations, to coordinate gross movements and muscle groups (such as torso and eyebrow) and enact top-down control of posture and internal

physiology, while preparing for and implementing actions. These proposed functions converge with the concept of allostatic regulation by which the brain anticipates upcoming changes in physiological demands on the basis of planned actions and exerts top-down preparatory control over the body[46].

## Human electrophysiology evidence

Penfield proposed the homunculus as an approximation of group-averaged, intraoperative direct electrocortical stimulation data, which showed significant overlap across patients and body parts. He later described his artistic rendering of the homunculus as “an aid to memory […] a cartoon of representation in which scientific accuracy is impossible”[2]. Re-examination of extant human stimulation data raises doubts about the veracity of the homunculus in individuals[11] and reveals an equal or better fit with the integrate–isolate model. In some individuals, a distal-to-proximal concentric organization was documented for the upper limb, just as in non-human primates[47], whereas face movements could be elicited from areas dorsal to the hand representation[48]. In addition to focal movements, several other response types are routinely elicited with M1 stimulation, all of which can be better accounted for by whole-organism control regions. Individuals have reported the urge to move while being aware that they are holding still; they have reported a sense of moving even though no movement is detectable; or they have moved but denied having done so[2]—effects consistent with modulation of a system also representing action goals. These responses are similar to those typically elicited in CON regions such as dACC[49] and anterior parietal cortex[50].

Stimulations almost never produce isolated torso or shoulder movements[47], and a common outcome of stimulation is no reported response at all[2]. Historically, stimulations that did not elicit movement were not documented. However, we re-analysed motor stimulations from a recent large study[51] by mapping them onto cortex, revealing a region that never elicited movement in any individual, corresponding to the middle inter-effector region (Extended Data Fig. 10a). These results suggest that stimulation strengths deemed safe in humans may not typically elicit movements in the M1 SCAN regions, akin to higher-order lateral and medial premotor regions[52].

Human brain–computer interface (BCI) recordings in M1, near the superior SCAN node, have also demonstrated whole-body movement tuning[53], possibly reflecting inter-effector activity and suggesting that the inter-effector motif could provide a target for whole-body BCI. Human speech BCI studies have suggested that the precentral gyrus between the hand and mouth effectors is essential for phonological-motoric aspects of speech planning[54], and that speech can also be decoded from a region at the bottom of precentral gyrus[55]. Following our identification of the SCAN, human depth electrode recordings verified that a portion of M1 between the foot and hand effector regions (superior SCAN) is active during foot, hand and mouth movements, strongly supporting its integrative, whole-body function[44].

# Evidence from clinical neurology

Brain lesion data further support the existence of dual systems for movement isolation and action integration, with partial redundancy in M1. Motor deficits after middle cerebral artery strokes are unilateral, more severe in most distal effectors, and without significant global organismal control deficits[56]. By contrast, lesions of SCAN-linked CON regions (dACC, anterior insula and aPFC) can cause isolated volitional deficits ranging from decreased fluency to abulia to akinetic mutism, with preserved motor abilities but little self-generated movement[57]. Similarly, anterior motor lesions in macaques can spare visually guided movements while selectively disrupting internally generated actions[58], whereas posterior lesions preserve intentionality but disrupt execution[59]. Animals with lesions in effector M1 typically recover gross effector control very quickly[60], whereas fine finger movement deficits persist longer[59,61]. More rapid recovery of gross motor abilities may be in part caused by proximal functions being taken up by the contra-lesional SCAN circuits, enabled by their bilateral spinal cord connectivity. Persistent deficits may therefore be more likely in functions uniquely supported by the effector-specific circuitry.

In an individual with extensive bilateral perinatal strokes but typical motor ability, extensive post-stroke reorganization maintained the SCAN patterning at the cost of part of the already reduced M1 hand area[32]. The top

third of M1 was destroyed, and surviving cortex contained an M1 hand area that was ventrally shifted and much smaller than in typical control brains. Surprisingly, SCAN regions were identified both above and below the surviving effector-specific hand region (Extended Data Fig. 2f), highlighting the importance of the SCAN for typical motor ability.

With specific connections to thalamic motor nuclei used as targets for clinical intervention (VIM and CM), the CON-linked SCAN may be relevant for a variety of movement disorders, including dystonia or essential tremor (Supplementary Information). Of particular note, many symptoms of Parkinson's disease span motor, physiological and volitional domains (for example, postural instability, autonomic dysfunction and reduced self-initiated activity, among many others[62]), mirroring SCAN connections to regions relevant for postural control (cerebellar vermis), volition and physiological regulation (CON[6,7,57]).

# Similarities to sensory systems

Many of the organizational features of M1 described here have clear parallels in sensory systems. Similar to the concentric somatotopic organization with fine finger movements at the centre, primary visual cortex over-represents higher acuity processing at the centre, concentrically transitioning to lower acuity in the periphery[63]. Similar to our integrate–isolate dual-systems model, visual processing streams are parallel and separated in thalamus, early visual cortex and higher-order visual processing streams, with each level of processing maintaining segregation of different types of information (for example, early: eccentricity versus angle[64]; late: faces versus objects[65]). Auditory processing may have similar features, as acoustic signals are processed at least partially in parallel for hearing and speech perception in superior temporal gyrus[66]. These findings suggest shared organizational principles across the brain's input and output processing streams. It is possible that S1 also includes concentric organizational elements, which should be explored in future work.

# A network for mind–body integration

Two behavioural control systems are interleaved in human M1. One well-known system consists of effector-specific circuits for precise, isolated movements of highly specialized appendages—fingers, toes and tongue—the type of dexterous motion needed for speaking or manipulating objects. A second, integrative output system, the SCAN, is more important for controlling the organism as a whole. The SCAN integrates body control (motor and autonomic) and action planning, consistent with the idea that aspects of higher-level executive control might derive from movement coordination[67]. The SCAN includes specific regions of M1, SMA, thalamus (VIM and CM), posterior putamen and the postural cerebellum, and is functionally connected to dACC regions linked to free will[68], parietal regions representing movement intentions[50], and insular regions for processing somatosensory, pain[9] and interoceptive visceral signals[33]. The apparent relative expansion of SCAN regions in humans could suggest a role in complex actions specific to humans, such as coordinating breathing for speech, and integrating hand, body and eye movement for tool use. A common factor across this wide range of processes is that they must be integrated if an organism is to achieve its goals through movement while avoiding injury and maintaining physiological allostasis[46]. The SCAN provides a substrate for this integration, enabling pre-action anticipatory postural, breathing, cardiovascular and arousal changes (such as shoulder tension, increased heart rate or 'butterflies in the stomach'). The finding that action and body control are melded in a common circuit could help explain why mind and body states so often interact.

# Methods

## Washington University adult participants

Data were collected from three healthy, right-handed, adult participants (aged 35, 25 and 27 years; 1 female) as part of a study investigating effects of arm immobilization on brain plasticity (previously published data[29,30,69]). Written informed consent was obtained from all participants. The study was approved by the Washington University School of Medicine Human Studies Committee and Institutional Review Board. The primary data employed here were collected either prior to the immobilization intervention (participants 1

and 3) or two years afterwards (participant 2). Data collected immediately after the intervention are presented for within-participant replication in Extended Data Fig. 1b. For details concerning data acquisition and processing, see ref. [29].

For two participants (participants 1 and 2), we collected additional fMRI data using the same sequence during performance of two motor tasks: a somatotopic mapping task and a motor control task.

### Movement task battery

A block design was adapted from the motor task in ref. [31]. In each run, the participant was presented with visual cues that directed them to perform one of five specific movements. Each block started with a 2.2-s cue indicating which movement was to be made. After this cue, a centrally presented caret replaced the instruction and flickered once every 1.1 s (without temporal jittering). Each time the caret flickered, participants executed the proper movement. Twelve movements were made per block. Each block lasted 15.4 s, and each task run consisted of 2 blocks of each type of movement as well as 3 blocks of resting fixation. Movements conducted within each run were as follows:

Run type 1: Close left (L) hand/Close right (R) hand/Flex L foot/Move tongue L and R (participant 1: 24 runs; participant 2: 20 runs).

Run type 2: Flex L elbow/Flex R elbow/Flex L wrist/Flex R wrist/Lift bilateral shoulders (participant 1: 10 runs; participant 2: 11 runs).

Run type 3: Flex L gluteus/Flex R gluteus/Tense abdomen/Open and close mouth/Swallow (participant 1: 10 runs; participant 2: 11 runs).

Run type 4: Flex L ankle/Flex R ankle/Bend L knee/Bend R knee/Flex bilateral toes (participant 1: 10 runs; participant 2: 11 runs).

Run type 5: Lift L eyebrow/Lift R eyebrow/Wink L eyelid/Wink R eyelid/Flare nostrils (participant 1: 10 runs; participant 2: 11 runs).

### Action control and coordination task

An event-related design implemented using JSpsych toolbox v6.3 was used to discriminate planning and execution of limb movement. See Supplementary Fig. 6 for an illustration of this task. Within the run, the participant is prompted to move either a single limb or to simultaneously move two limbs. There are four possible motions—open–close of fingers or toes, left–right flexion of the wrist or ankle, clockwise rotation of the wrist or ankle, and anticlockwise rotation of the wrist or ankle—each of which may be executed by any of the four extremities (left or right upper or lower extremity). Each motion–extremity combination may be required in isolation, or in combination with a second simultaneous motion. The participant is cued to prepare the movement(s) when they see one or two movement symbols placed on a body shape in a grey colour (planning phase), and is then cued to execute the movement(s) when the grey symbol or symbols turn green (execution phase). Using a pseudorandom jitter, the planning phase can last from 2 to 6.5 s followed by 4 to 8.5 s of movement execution. Each movement trial (planning and execution) is followed by a jittered fixation of up to 5 s. A rest block of 8.6 s is implemented every 12 movements. Two possible movements are requested during the task run and practiced before the task. The movement pair is changed for each task run. 48 trials were collected in each run. Twelve total runs were acquired per participant.

**Laryngeal mapping task**

For the same two participants, as well as for one additional participant (participant 8, 40 years of age, from whom written informed consent was obtained) additional fMRI data were collected during performance of a laryngeal mapping task using a multiband five-echo blood oxygen level-dependent (BOLD) contrast sensitive gradient echo-planar sequence (flip angle = 68°, resolution = 2.0 mm isotropic, TR = 1,761 ms, multiband 6 acceleration, $TE_1$: 14.20 ms, $TE_2$: 38.93 ms, $TE_3$: 63.66 ms, $TE_4$: 88.39 ms, and $TE_5$: 113.12 ms), with each run lasting 3 min 52 s. A pair of spin-echo echo-planar images (EPI) with opposite phase-encoding directions (anterior→posterior (AP) and posterior→anterior (PA)) but identical geometrical parameters to the BOLD sequence were acquired. In participant 8, 15 min of resting-state fMRI were also acquired using the sequence above.

An additional set of ten movement task runs adapted from[31] were collected to localize laryngeal phonation. In each run, the participant was presented with visual cues that directed them to perform one of six specific movements: Left hand, Right hand, Left foot, Right foot, Tongue or Voice. In the Voice condition, participants were required to briefly make the noise "eeee" without moving their jaw. Each block started with a 3.0 s cue indicating which movement was to be made. After this cue, a centrally presented caret replaced the instruction and flickered once every 1.0 s (without temporal jittering). Each time the caret flickered, participants executed the proper movement. Ten movements were made per block. Each block lasted 15.0 s, and each task run consisted of 2 blocks of each type of movement as well as 2 blocks of resting fixation. Each participant completed 10 runs.

## Cornell adult participants

Data were collected from four healthy adult participants (ages 29, 38, 24 and 31; all male) as part of a previously published study[70]. The study was approved by the Weill Cornell Medicine Institutional Review Board. Written informed consent was provided by each participant.

For details concerning data acquisition and processing, see ref. [70].

## Neonatal participant

Data were collected from one sleeping, healthy full-term neonatal participant beginning 13 days after birth, corresponding to 42 weeks post-menstrual age. The study was approved by the Washington University School of Medicine Human Studies Committee and Institutional Review Board. Written informed consent was provided by a parent.

### MRI acquisition

The participant was scanned while asleep over the course of 4 consecutive days using a Siemens Prisma 3T scanner on the Washington University Medical Campus. Every session included collection of a high-resolution T2-weighted spin-echo image (TE = 563 ms, TR = 3,200 ms, flip angle = 120°,

208 slices with 0.8 × 0.8 × 0.8 mm voxels). In each session, a number of 6 min 45 s multi-echo resting-state fMRI runs were collected as a five-echo BOLD contrast sensitive gradient echo-planar sequence (flip angle = 68°, resolution = 2.0 mm isotropic, TR = 1,761 ms, multiband 6 acceleration, $TE_1$: 14.20 ms, $TE_2$: 38.93 ms, $TE_3$: 63.66 ms, $TE_4$: 88.39 ms, and $TE_5$: 113.12 ms). The number of BOLD runs collected in each session depended on the ability of the neonate to stay asleep during that scan; across the 4 days, 23 runs were collected in total. A pair of spin-echo EPI images with opposite phase-encoding directions (AP and PA) but identical geometrical parameters and echo spacing were acquired between every three BOLD runs or any time the participant was removed from the scanner.

### MRI processing

Structural and functional processing followed the pipeline used for the Washington University dataset, with two exceptions. First, segmentation, surface delineation and atlas registration were conducted using a T2-weighted image (the single highest quality T2 image, as assessed via visual inspection) rather than a T1-weighted image, due to the inverted image contrast observed in neonates. Second, after the multi-echo BOLD data were unwarped and normalized to atlas space, it was optimally combined before nuisance regression and mapping to cifti space. All fMRI scans from the second day of scanning were excluded due to registration abnormalities.

## Infant participant

Data were collected from one healthy sleeping infant aged 11 months. The study was approved by the Washington University School of Medicine Human Studies Committee and Institutional Review Board. Written informed consent was provided by a parent.

### MRI acquisition

The participant was scanned while asleep over the course of three sessions using a Siemens Prisma 3T scanner on the Washington University Medical Campus. The first session included collection of a high-resolution T1-

weighted MP-RAGE (TE = 2.24 ms, TR = 2,400 ms, flip angle = 8°, 208 slices with 0.8 × 0.8 × 0.8 mm voxels) and a T2-weighted spin-echo image (TE = 564 ms, TR = 3200 ms, flip angle = 120°, 208 slices with 0.8 × 0.8 × 0.8 mm voxels). The second and third sessions included collection of 26 total runs of resting-state fMRI, each collected as a 6 min 49 s-long BOLD contrast sensitive gradient echo-planar sequence (flip angle = 52°, resolution = 3.0 mm isotropic, TE = 30 ms, TR = 861 ms, multiband 4 acceleration). For each run, a pair of spin-echo EPI images with opposite phase-encoding directions (AP and PA) but identical geometrical parameters and echo spacing were acquired to correct spatial distortions.

### MRI processing

Structural processing followed the DCAN Labs processing pipeline found in the ABCD BIDS Community Collection (ABCC; NDA Collection 3165) (https://github.com/DCAN-Labs/abcd-hcp-pipeline)[71], which we found performed the best surface segmentation at this age. Functional processing followed the pipeline used for the Washington University adult dataset.

## Child participant

Data were collected from one healthy awake male child age 9 years. The study was approved by the Washington University School of Medicine Human Studies Committee and Institutional Review Board. Written informed consent was provided by a parent and assent was given by the participant.

### MRI acquisition

The participant was scanned repeatedly over the course of 12 sessions using a Siemens Prisma 3T scanner on the Washington University Medical Campus. These sessions included collection of 14 high-resolution T1-weighted MP-RAGE images (TE = 2.90 ms, TR = 2,500 ms, flip angle = 8°, 176 slices with 1 mm isotropic voxels), 14 T2-weighted spin-echo images (TE = 564 ms, TR = 3200 ms, flip angle = 120°, 176 slices with 1 mm isotropic voxels), and 26 total runs of resting-state fMRI, each collected as a 10 min-long BOLD contrast sensitive gradient echo-planar sequence (flip

angle = 84°, resolution = 2.6mm isotropic, 56 slices, TE = 33 ms, TR = 1,100 ms, multiband 4 acceleration). In each session, a pair of spin-echo EPI images with opposite phase-encoding directions (AP and PA) but identical geometrical parameters and echo spacing were acquired to correct spatial distortions in the BOLD data.

**MRI processing**

Structural and functional processing followed the DCAN Labs processing pipeline found in the ABCD BIDS Community Collection (ABCC; NDA Collection 3165)[71] (https://github.com/DCAN-Labs/abcd-hcp-pipeline).

## Participant with perinatal stroke

PS1, a left-handed, 13-year-old male who played for a competitive youth baseball team, was referred to an orthopaedic physician because of difficulty using his right arm effectively. Ulnar neuropathy was considered and he was referred for physical therapy. However, PS1 was first seen by a child neurologist (N.U.F.D.) for further evaluation. Structural brain MRI revealed unexpectedly extensive bilateral cystic lesions consistent with perinatal infarcts. Review of PS1's medical history revealed that the injury occurred in the perinatal period.

Data acquisition from PS1 were performed with the approval of the Washington University Institutional Review Board. Written informed consent was provided by PS1's mother and assent was given by PS1 at the time of data acquisition.

For additional details regarding clinical history, neuropsychological evaluations, motor assessments, or MR image acquisition or processing, see ref. [32].

## UMN macaque

Data were collected from a sedated adult female macaque monkey (*Macaca fascicularis*) aged 6 years. Experimental procedures were carried out in accordance with the University of Minnesota Institutional Animal Care and

Use Committee and the National Institute of Health standards for the care and use of non-human primates. The subject was fed ad libitum and pair-housed within a light- and temperature-controlled colony room. The animal was not water restricted. The subject did not have any prior implant or cranial surgery. The animal was fasted for 14–16 h prior to imaging. On scanning days, anaesthesia was first induced by intramuscular injection of atropine (0.5 mg $kg^{-1}$), ketamine hydrochloride (7.5 mg $kg^{-1}$), and dexmedetomidine (13 µg $kg^{-1}$). The subject was transported to the scanner anteroom and intubated using an endotracheal tube. Initial anaesthesia was maintained using 1.0%–2% isoflurane mixed with oxygen (1 l $min^{-1}$ during intubation and 2 l $m^{-1}$ during scanning to compensate for the 12-m length of the tubing used). For functional imaging, the isoflurane level was lowered to 1%. The subject was placed onto a custom-built coil bed with integrated head fixation by placing stereotactic ear bars into the ear canals. The position of the animal corresponds to the sphinx position. Experiments were performed with the animal freely breathing. Continuous administration of 4.5 µg $kg^{-1}$ $h^{-1}$ dexmedetomidine using a syringe pump was administered during the procedure. Rectal temperature (~37.6 °C), respiration (10–15 breaths per min), end-tidal $CO_2$ (25–40), electro-cardiogram (70–150 bpm), and peripheral capillary oxygen saturation ($SpO_2$) (>90%) were monitored using an MRI compatible monitor (IRAD-IMED 3880 MRI Monitor). Temperature was maintained using a circulating water bath as well as chemical heating pads and padding for thermal insulation.

**MRI acquisition**

Data were acquired on a Siemens Magnetom 10.5 T Plus. A custom in-house built and designed RF coil was used with an 8-channel transmit/receive end-loaded dipole array of 18-cm length combined with a close-fitting 16-channel loop receive array head cap, and an 8-channel loop receive array of 50 × 100 $mm^2$ size located under the chin[72]. A B1+ (transmit B1) field map was acquired using a vendor provided flip angle mapping sequence and then power calibrated for each individual. Following B1+ transmit calibration, 3–5 averages (23 min) of a T1-weighted MP-RAGE were acquired for anatomical processing (TR = 3300 ms, TE = 3.56 ms, TI = 1,140, flip angle = 5°, slices = 256, matrix = 320×260, acquisition voxel size = 0.5 × 0.5 × 0.5

$mm^3$, in-plane acceleration GRAPPA = 2). A resolution and field of view-matched T2-weighted 3D turbo spin-echo sequence was run to facilitate B1 inhomogeneity correction. Five images were acquired in both phase-encoding directions (R→L and L→R) for offline EPI distortion correction. Six runs of fMRI time series, each consisting of 700 continuous 2D multiband EPI[73,74,75] functional volumes (TR = 1,110ms; TE = 17.6 ms; flip angle = 60°, slices = 58, matrix = 108 × 154; field of view = 81 × 115.5 mm ; acquisition voxel size = 0.75 × 0.75 × 0.75 mm) were acquired with a left–right phase-encoding direction using in-plane acceleration factor GRAPPA = 3, partial Fourier = 7/8, and MB factor = 2. Since the macaque was scanned in sphinx position, the orientations noted here are what is consistent with a (head first supine) typical human brain study (in terms of gradients) but translate differently to the actual macaque orientation.

### MRI processing

Processing followed the DCAN Labs non-human primate processing pipeline (http://github.com/DCAN-Labs/nhp-abcd-bids-pipeline), with minor modifications. Specifically, we observed that distortion from the 10T scanner was so extensive that the field maps did not fully correct it. Therefore, instead of field map-based unwarping, we used the computed field map-based warp as an initial starting point for Synth, a field map-less distortion correction algorithm that creates synthetic undistorted BOLD images for registration to anatomical images[76]. Synth substantially reduced residual BOLD image distortion.

## PRIME-DE macaque

Raw structural and functional data were provided from the Oxford dataset of the PRIMatE Data Exchange (PRIME-DE) consortium (https://fcon_1000.projects.nitrc.org/indi/PRIME/oxford.html)[77,78]. The full dataset consisted of 19 (age 4.1 ± 0.98 years, weight 6.61 ± 2.94 kg) rhesus macaques (*Macaca mulatta*). The animal care, anaesthesia and MRI protocols were carried out in accordance with the UK Animals (Scientific Procedures) Act of 1986. Animals in the study were group-housed prior to scanning. Ketamine (10 mg $kg^{-1}$) was administered via intramuscular

injection for induction, along with either xylazine (0.125–0.25 mg kg$^{-1}$), midazolam (0.1 mg kg$^{-1}$), or buprenorphine (0.01 mg kg$^{-1}$). Additionally, injections of atropine (0.05 mg kg$^{-1}$, intramuscular injection), meloxicam (0.2 mg kg$^{-1}$, intravenous injection), and ranitidine (0.05 mg kg$^{-1}$) were administered. A minimum of 15 min prior to being placed in the stereotaxic frame, animals also received local anaesthetics (5% lidocaine/prilocaine cream and 2.5% bupivacaine injected subcutaneously around ears). Finally, anaesthesia was maintained with isoflurane, and scanning began 1.5–2 h after the initial ketamine induction.

### MRI acquisition

Anaesthetized animals were placed in the sphinx position into a stereotactic frame (Crist Instrument) and scanned in a horizontal 3T MRI scanner using a four-channel phased-array coil (Windmiller Kolster Scientific, Fresno, CA). Each animal received 53.33 min (1,600 volumes) of resting-state data, which was acquired at a 2.0 mm isotropic voxel resolution (TR = 2,000 ms, TE = 19 ms, Flip angle = 90°). A T1-weighted MP-RAGE sequence was used to acquire anatomical data (TR = 2,500 ms, TE = 4.01 ms, TI = 1,100, flip angle = 8°, acquisition voxel size = 0.5 × 0.5 × 0.5 mm, 128 slices).

### MRI processing

Processing for structural data followed the DCAN Labs non-human primate processing pipeline (https://github.com/DCAN-Labs/nhp-abcd-bids-pipeline). Smoothing was applied with FWHM = 1.5 mm in both volume and surface space. The surface data were then down-sampled to a 10k surface to create the preprocessed cifti data. Finally, each animal's data was closely visually inspected for quality. Following these inspections, data from 11 animals were excluded due to the presence of artefact in or near the central sulcus, leaving eight animals in the final data. This sample size of eight was chosen to include all available artefact-free data. No randomization or blinding was performed.

## Group-averaged datasets

Resting-state fMRI data was averaged across participants within each of five large datasets.

**UK Biobank**

A group-averaged weighted eigenvectors file from an initial batch of 4,100 UKB participants aged 40–69 years (53% female) scanned using resting-state fMRI for 6 min was downloaded from https://www.fmrib.ox.ac.uk/ukbiobank/. This file consisted of the top 1,200 weighted spatial eigenvectors from a group-averaged principal component analysis. See ref. [79] and documentation at https://biobank.ctsu.ox.ac.uk/crystal/ukb/docs/brain_mri.pdf for details of the acquisition and processing pipeline. This eigenvectors file was mapped to the Conte69 surface template atlas[80] using the ribbon-constrained method in Connectome Workbench[81], and the eigenvector time courses of all surface vertices were cross-correlated.

**Adolescent Brain Cognitive Development Study**

Twenty minutes (4 × 5-min runs) of resting-state fMRI data, as well as high-resolution T1-weighted and T2-weighted images, were collected from 3,928 9- to 10-year-old participants (51% female), who were selected as the participants with at least 8 min of low-motion data from a larger scanning sample. Data collection was performed across 21 sites within the USA, harmonized across Siemens, Philips and GE 3T MRI scanners. See ref. [82] for details of the acquisition parameters. Data processing was conducted using the ABCD-BIDS pipeline found in the ABCD BIDS Community Collection (ABCC; NDA Collection 3165) (https://github.com/DCAN-Labs/abcd-hcp-pipeline)[71]; see ref. [83] for details.

**Human Connectome Project**

A vertexwise group-averaged functional connectivity matrix from the HCP 1200 participants release was downloaded from https://db.humanconnectome.org. This matrix consisted of the average strength of functional connectivity across all 812 participants aged 22–35

years (410 female) who completed 4× 15-min resting-state fMRI runs and who had their raw data reconstructed using the newer recon 2 software. See refs. [81,84,85,86] for details of the acquisition and processing pipeline.

### Washington University 120

Data were collected from 120 healthy young adult participants recruited from the Washington University community during relaxed eyes-open fixation (60 females, ages 19–32). Scanning was conducted using a Siemens TRIO 3.0T scanner and included collection of high-resolution T1-weighted and T2-weighted images, as well as an average of 14 min of resting-state fMRI. See ref. [87] for details of the acquisition and processing pipeline.

### Neonates

Mothers were recruited during the second or third trimester from two obstetrics clinics at Washington University as part of the Early Life Adversity, Biological Embedding, and Risk for Developmental Precursors of Mental Disorders (eLABE) study. This study was approved by the Human Studies Committees at Washington University in St. Louis and written informed consent was obtained from mothers. Neuroimaging was conducted in full-term, healthy neonate offspring shortly after birth (average post-menstrual age of included participants 41.4 weeks, range 38–45 weeks). Of the 385 participants scanned for eLABE, 262 were included in the current analyses (121 female). See ref. [88] for additional details of the participants, criteria for exclusion, scanning acquisition protocol and parameters, and processing pipeline.

## Analyses

### Functional connectivity

For each single-participant dataset, a vertex or voxelwise functional connectivity matrix was calculated from the resting-state fMRI data as the Fisher-transformed pairwise correlation of the time series of all vertices/voxels in the brain. In the ABCD, Washington University 120,

eLABE and PRIME-DE datasets, vertex and voxelwise group-averaged functional connectivity matrices were constructed by first calculating the vertex or voxelwise functional connectivity within each participant as the Fisher-transformed pairwise correlation of the time series of all vertices or voxels in the brain, and then averaging these values across participants at each vertex or voxel.

## Seed-based functional connectivity

We defined a continuous line of seeds down the left precentral gyrus by selecting every vertex in a continuous straight line on the cortical surface between the most ventral aspect of the medial motor area (approximate MNI coordinates (−4, −31, 54)) and the ventral lip of the precentral gyrus right above the operculum (approximate MNI coordinates (−58, 4, 8)). For each seed, we examined its map of functional connectivity as the Fisher-transformed correlation between that vertex's time course and that of every other vertex or voxel in the brain.

## Network detection in somatomotor cortex

To define the somatomotor regions that were visually identified from the seed-based connectivity analysis in an unbiased fashion for further exploration, we entered each individual adult human participant's data into a data-driven network detection algorithm designed to identify network subdivisions that are hierarchically below the level of classic large-scale networks (for example, those that produce hand/foot divisions in somatomotor cortex;[23,37]). We have previously described how this approach identifies sub-network structures that converge with task-activated regions[89] and with known neuroanatomical systems[90].

In each adult participant, this analysis clearly identified network structures corresponding to motor representation of the foot, hand and mouth; and it additionally identified network structures corresponding exactly to the previously unknown connectivity pattern identified from the seed-based connectivity exploration as the inter-effector regions. For simplicity, we manually grouped all inter-effector subnetworks together as a single putative network structure (labelled as inter-effector) for further analysis.

Finally, to identify classic large-scale networks in each participant, we repeated the Infomap algorithm on matrices thresholded at a series of denser thresholds (ranging from 0.2% to 5%), and additionally identified individual-specific networks corresponding to the default, medial and lateral visual, cingulo-opercular, fronto-parietal, dorsal attention, language, salience, parietal memory, and contextual association networks following procedures described in ref. [24]. See Supplementary Fig. 6 for these individual-specific networks.

## Differences in functional connectivity between inter-effector and foot, hand or mouth regions

Within each adult human participant, we calculated an inter-effector connectivity map as the Fisher-transformed correlation between the average time course of all cortical inter-effector vertices and the time course of every other vertex or voxel in the brain. We then repeated this procedure to calculate a connectivity map for the foot, hand and mouth areas.

To identify brain regions more strongly connected to inter-effector regions than to other motor regions, we computed the smallest positive difference in each voxel or vertex between inter-effector connectivity and any foot, hand or mouth connectivity. That is, we calculated (inter-effector – max[foot, hand, mouth]). This represents a conservative approach that only identifies regions of the brain for which the inter-effector regions are more strongly connected than any of the other motor areas.

## Differences in functional connectivity among inter-effector regions

Within each adult human participant, as well as in the HCP group-averaged data, we computed a connectivity map for each of the three distinct inter-effector regions (superior, middle and inferior) as the Fisher-transformed correlation between the average time course of all cortical vertices in the two bilateral regions in each position and the time course of every other vertex or voxel in the brain.

To identify brain regions more strongly connected to one of the inter-effector regions than the other two, we computed the smallest positive difference in each voxel or vertex between that region's connectivity and either of the other two regions' connectivity. That is, we calculated (superior inter-effector – max[middle, inferior inter-effector]), (middle inter-effector – max[superior, inferior inter-effector]), and (inferior inter-effector – max[superior, middle inter-effector]). This represents a conservative approach that only identifies regions of the brain for which one inter-effector region is more strongly connected than either of the other two regions.

## Functional connectivity with CON

Within each adult human participant, we calculated the functional connectivity between each of the foot, hand, mouth, and inter-effector regions and the CON. This was computed as the Fisher-transformed correlation between (1) the average time course across all vertices in the motor region and (2) the average time course across all vertices in the CON. We conducted paired *t*-tests across subjects comparing the inter-effector connectivity with CON against each of the foot, hand and mouth connectivity strengths, FDR-correcting for the three tests conducted.

We then calculated the functional connectivity between the inter-effector regions and every other large-scale cortical network in the brain (visual, auditory, salience, premotor, fronto-parietal, default mode, dorsal attention, language, contextual association, and parietal memory). The strength of connectivity between the inter-effector network and the CON was compared against the strength of its connectivity to each of these other networks using paired *t*-tests, FDR-correcting for the ten tests conducted.

## Motor and CON network visualization

Visualization of network relationships was conducted using spring-embedded plots[23], as implemented in Gephi (https://gephi.org/). In each individual adult human participant, nodes were defined as congruent clusters of foot, hand, mouth, inter-effector, and CON networks larger than 20 $mm^2$. Pairwise connectivity between nodes was calculated as the Fisher-transformed correlation of their mean time courses. For visualization

purposes, graphs were constructed by thresholding the pairwise node-to-node connectivity matrices at 40% density (the general appearance of the graphs did not change across a range of densities).

## Functional connectivity with adjacent postcentral gyrus

In each adult human participant, we defined the pre- and postcentral gyri based on the individual-specific Brodmann areal parcellation produced by Freesurfer, which was deformed into fs_LR_32k space to match the functional data. Precentral gyrus was considered to be the vertices labelled as BA 4a and 4p, and postcentral gyrus was the vertices labelled as BA 3b and 2. BA 3a (fundus of central sulcus) was not considered for this analysis. Because the medial aspect of somatomotor cortex (corresponding to representation of the leg and foot) was always classified by Freesurfer as BA 4a, we defined the medial postcentral gyrus as the cortical vertices with *y*-coordinates farther posterior than the median *y*-coordinate of the foot region (from the network mapping above).

Within the participant's precentral gyrus, we labelled vertices as representing foot, hand, mouth or inter-effector according to their labels from the network mapping procedure. We then partitioned the postcentral gyrus into foot, hand, mouth and inter-effector areas depending on which precentral region each vertex was physically closest to. Finally, within each partition (foot, hand, mouth and inter-effector) we calculated the average connectivity between the pre and postcentral gyrus as the Fisher-transformed correlation between the average time courses of all vertices in each area. We then conducted paired *t*-tests across subjects comparing the inter-effector connectivity with adjacent S1 against each of the foot/hand/mouth connectivity strengths with S1, FDR-correcting for the three tests conducted.

## Functional connectivity with middle insula

In each adult human participant, we defined the middle insula based on the individual-specific Freesurfer gyral parcellation using the Destrieux atlas[91], which was deformed into fs_LR_32k space to match the functional data. Middle insula was considered to be the vertices labelled as the superior segment of the circular sulcus of the insula or as the short insular gyrus. We then calculated the functional connectivity between each of the bilateral foot, hand, mouth, and inter-effector regions and the bilateral middle insula.

We conducted paired *t*-tests across subjects comparing the inter-effector connectivity with middle insula against each of the foot, hand and mouth connectivity strengths, FDR-correcting for the number of tests conducted.

## Functional connectivity with cerebellum

In each adult human participant, we calculated the functional connectivity between each of the foot, hand, mouth and inter-effector regions with each voxel of the cerebellum. Cerebellar connectivity strengths calculated this way were then mapped onto a cerebellar flat map using the SUIT toolbox[92]. Connectivity strengths were averaged within each of 27 atlas regions[93]. For each region, we conducted three paired t-tests comparing inter-effector connectivity strength against foot, hand and mouth connectivity strength, FDR-correcting for the total number of tests conducted. Regions were reported if the inter-effector connectivity strength was significantly higher than the connectivity strength of all other motor regions.

## Functional connectivity with putamen

In each adult human participant, we divided each unilateral putamen in each hemisphere into quarters by splitting it based on the median of its *y* (anterior-posterior) and *z* (dorsal–ventral) coordinates. We then calculated the functional connectivity between each of the foot, hand, mouth and inter-effector regions and each putamen quarter.

For each putamen division, we conducted paired *t*-tests across subjects comparing the inter-effector connectivity with that putamen division against each of the foot, hand and mouth connectivity strengths, FDR-correcting for the number of tests conducted. We reported divisions in which the inter-effector connectivity was significantly different from all three effector-specific connectivities.

## Functional connectivity with thalamus

To investigate subregions of thalamus, we employed the DISTAL atlas v1.1[94], which contains a number of histological thalamic subregions

identified by[95]. This atlas was down-sampled into the 2-mm isotropic space of the functional data. Functional connectivity maps seeded from the foot, hand, mouth, and inter-effector regions in each adult human participant were computed, and mean connectivity values were calculated within each atlas region. The atlas specifies multiple subregions for many nuclei; these subregions were combined and treated as single nuclei for the purposes of connectivity calculation.

For each adult human participant, we averaged the connectivity seeded from the inter-effector regions and from each of the foot, hand and mouth regions across all voxels within each thalamic nucleus. For each thalamic nucleus, we conducted paired *t*-tests across subjects comparing the inter-effector with the mean of the foot, hand and mouth connectivity strengths, FDR-correcting for the number of thalamic nuclei tested.

## Lag structure of RSFC

We used a previously published method for estimating relative time delays (lags) in fMRI data[34,96]. In brief, for each session in each adult human participant, we computed a lagged cross-covariance function (CCF) between each pair of vertex or voxel time courses within the motor system and CON in the cortex. Lags were more precisely determined by estimating the cross-covariance extremum of the session-level CCF using three-point parabolic interpolation. The resulting set of lags was assembled into an antisymmetric matrix capturing all possible pairwise time delays (TD matrix) for each session, which was averaged across sessions to yield participant-level TD matrices. Finally, each participant's TD matrix was averaged across rows to summarize the average time-shift from one vertex to all other vertices. Average time lag was then averaged across all vertices with each of the precentral gyrus foot, hand, mouth and inter-effector regions, and the CON.

We then conducted paired *t*-tests across subjects comparing (1) the mean lag in inter-effector regions against the mean lags in each of the foot, hand and mouth regions, and (2) the mean lag in CON regions against the mean lags in each of the foot, hand and mouth regions.

## Macaque RSFC

We placed connectivity seeds continuously along area 4p in the left hemisphere of each macaque, as well as continuously running from the dorsal cingulate motor area to the rostral cingulate motor area in the dACC (area 24). See Supplementary Fig. 8a for medial cortex seed locations and Supplementary Fig. 8b for all functional connectivity maps from all medial seeds.

### Structural MRI

## Cortical thickness

Within each adult human participant, the map of cortical thickness generated by the Freesurfer segmentation was deformed into fs_LR_32k space to match the functional data. Precentral gyrus foot, hand, mouth and inter-effector regions were defined as above, and mean cortical thickness was calculated within each region. We then conducted paired *t*-tests across subjects comparing the inter-effector thickness against each of the foot, hand and mouth thicknesses, correcting for the three tests conducted.

## Fractional anisotropy

White matter fibres tracked from separate areas of M1 using diffusion imaging quickly converge into the internal capsule and become difficult to dissociate. As such, we tested for fractional anisotropy differences in the white matter immediately below the precentral gyrus.

To calculate fractional anisotropy beneath the cortex, we first constructed fs_LR_32k-space surfaces 2 mm below each grey-white surface in adult human participants 1–3. To accomplish this, for each vertex on the surface, we computed the 3D vector between corresponding points on the fs_LR_32k pial and the grey-white surfaces, and we extended that vector an additional 2 mm beyond the grey-white surface in order to create a lower surface. We then mapped the fractional anisotropy values using the using the ribbon-constrained method, mapping between the grey-white and the 2

mm-under surfaces. The result is fractional anisotropy values mapped to a lower surface within white matter that is in register to the existing fs_LR_32k surfaces on which the functional data is mapped and the motor regions defined.

Precentral gyrus foot, hand, mouth and inter-effector regions were defined as above, and we calculated mean fractional anisotropy beneath each cortical region.

We then conducted paired *t*-tests across subjects comparing the mean fractional anisotropy beneath the inter-effector regions against mean fractional anisotropy beneath each of the foot, hand and mouth regions.

## Myelin density

Within each adult human participant, we created vertexwise maps of intracortical myelin content following methods described in refs. [81,97]. Precentral gyrus was defined as above. Across participants, we found that baseline myelin density values (both in precentral gyrus and in the whole-brain myelin density map) varied wildly across participants in different datasets, likely based on differences in the T1- and T2-weighted sequences employed. Thus, for optimal visualization of results, in each participant we normalized the myelin density values by dividing the calculated vertexwise myelin densities in precentral gyrus by the mean myelin density across the whole precentral gyrus. Finally, precentral gyrus foot, hand, mouth and inter-effector regions were defined as above, and mean normalized myelin density was calculated within each region. We then conducted paired *t*-tests across subjects comparing the inter-effector myelin density against each of the foot, hand and mouth myelin densities, correcting for the three tests conducted.

### Task fMRI

## Movement task battery analysis

Basic analysis of the movement task battery data was conducted using within-participant block designs. To compute the overall degree of

activation in response to each motion, data from each run was entered into a first-level analysis within FSL's FEAT[98] in which each motion block was modelled as an event of duration 15.4 s, and the combined block waveform for each motion condition was convolved with a haemodynamic response function to form a separate regressor in a generalized linear model (GLM) analysis testing for the effect of the multiple condition regressors on the time course of activity within every vertex or voxel in the brain. Beta value maps for each condition were extracted for each run and entered into a second-level analysis, in which run-level condition betas were tested against a null hypothesis of zero activation in a one-sample *t*-test across runs (within participant). The resulting *t*-values from each motion condition tested in this second-level analysis were converted to *Z*-scores. *Z*-score activation maps were smoothed with a geodesic 2D (for surface data) or Euclidean 3D (for volumetric data) Gaussian kernel of $\sigma = 2.55$ mm.

## Movement task battery winner take all

For each vertex within the broad central sulcus area, we identified the movement that produced the greatest activation strength (*Z*-score from second-level analysis, above) in that vertex, and we assigned that motion to that vertex.

## Movement task battery curve fitting

For each vertex within precentral gyrus, we first computed its position along the dorsal–ventral axis of left hemisphere M1. This was done by identifying the closest point within the continuous line of points running down precentral gyrus (defined in 'Seed-based functional connectivity'), and assigning that closest point's ordered position within the line to the vertex.

For every movement, we then plotted that dorsal–ventral M1 position against *Z*-score activation in each vertex. We then fit two curves to each of these relationships. The first curve was a single-Gaussian model of the form:

$$\text{Activation} = a_1 \times \exp(-((\text{position-}b_1)/c_1)^2).$$

The second curve was a double-Gaussian model of the form:

Activation = $a_1 \times \exp(-((\text{position-}b_1)/c_1)^2) + a_2 \times \exp(-((\text{position-}b_2)/c_2)^2)$.

The $a_1$ and $a_2$ parameters in each model were constrained to be positive (to enforce positive-going peaks). Curve fitting was constrained to be conducted within the general vicinity of the activated area in order to avoid fitting negative activations observed in distant portions of M1. For lower extremity movements, this meant excluding the bottom third of M1; for upper extremity movements, the bottom third of M1 plus the medial wall; for face movements, the top third of M1.

Finally, we tested whether the one- or two-peak models better fit the data. This was done by conducting an *F*-test between the models, computed as:

$$F=(({{\rm{SSE}}}_{1{\rm{peak}}}-{{\rm{SSE}}}_{2{\rm{peaks}}})/({{\rm{df}}}_{1{\rm{peak}}}-{{\rm{df}}}_{2{\rm{peaks}}}))/({{\rm{SSE}}}_{2{\rm{peaks}}}/{{\rm{df}}}_{2{\rm{p}}{\rm{e}}{\rm{a}}{\rm{k}}{\rm{s}}}).$$

where SSE represents the sum of squared errors from the model and df represents the degrees of freedom in the model.

The $P$ value was computed from this $F$ by employing the $F$-statistic continuous distribution function (fcdf.m) in Matlab and using ($df_{1peak}$ – $df_{2peaks}$) and $df_{2peaks}$ as the numerator and denominator degrees of freedom, respectively.

## Movement task battery curve visualization

For each movement, the complete dorsal–ventral M1 position versus *Z*-score activation profile (from above) was visualized more clearly by fitting a LOWESS curve. These LOWESS curves recapitulated the two-peak activation fits while also revealing additional task responsive cortex.

## Movement selectivity

Based on results from the above winner-take-all analysis, we identified the movement that was most preferred at the centre of each the three effector-specific (toe movement, hand movement and tongue movement) and inter-effector regions (abdominal movement, eyelid movement and swallowing). The centre-most movements were selected to avoid issues with spreading, overlapping activation near the borders of effector-specific and inter-effector regions. For every vertex within the precentral gyrus, we compared the strength of activation between the most preferred of the six movements at that vertex against the activation of the second-most preferred movements. The differences between these activation strengths was taken to be the movement selectivity of that vertex.

## Movement coactivation

For each region among the six resting-state-derived foot, hand, mouth, and inter-effector regions in the precentral gyrus, we calculated the average activation within that region for each movement, producing a profile of motor activation strengths for that region. We also calculated the average activation within all CON vertices for each movement. To determine the degree to which various regions were coactive across movements, we then correlated each foot, hand, mouth and inter-effector cluster's profile of activation strengths with that of all other clusters, and with that of the CON. Note: visualization of activation maps revealed some striping, suggesting that the Open and close mouth and the Bend L knee conditions were partially distorted by head motion; therefore, these conditions were excluded from analysis, although their inclusion did not change results.

## Laryngeal motor mapping task analysis

As with the movement task battery, analysis of the laryngeal mapping task data was conducted using within-participant block designs. To compute the overall degree of activation in response to each motion, data from each run was entered into a first-level analysis within FSL's FEAT[98] in which each motion block was modelled as an event of duration 15.0 s, and the combined block waveform for each motion condition was convolved with a hemodynamic response function to form a separate regressor in a GLM analysis testing for the effect of the multiple condition regressors on the

time course of activity within every vertex or voxel in the brain. Beta value maps for each condition were extracted for each run and entered into a second-level analysis, in which run-level condition betas were tested against a null hypothesis of zero activation in a one-sample *t*-test across runs (within participant). The resulting *t*-values from each motion condition tested in this second-level analysis were converted to *Z*-scores.

Note that the laryngeal mapping data was not included in the movement task battery analysis because it was collected on a different scanner with a different sequence, and so would not be directly comparable.

## Action control and coordination task analyses

Analysis of the action control task was conducted using within-participant event-related designs. For each separate run, a GLM model was constructed in FEAT[98] in which separate regressors described the initiation of (1) planning and (2) execution of each type of movement (4 movements × 4 limbs). Each regressor was constructed as a 0-length event convolved with a canonical haemodynamic response, and beta values for each regressor were estimated for every voxel in the brain. These beta value maps for each condition were thus computed for each run and entered into a second-level analysis, in which a *t*-test across runs contrasted the run-level planning betas against the run-level execution betas.

### Human direct electrocortical stimulation site mapping

Each stimulation location reported in ref. [51] was separately mapped into the MNI-space Conte69 atlas pial cortical surface[80] by identifying the vertex with the minimal Euclidean distance to the stimulation site's MNI coordinates. Movements resulting from each site were classified as 'lower extremity', 'upper extremity' or 'face' and coloured accordingly (although no lower extremity movements were reported in the displayed left hemisphere).

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Data from individual subjects participants 1–3 are available in the openneuro repository: https://openneuro.org/datasets/ds002766/versions/3.0.0. Data from the individual perinatal stroke subject are available in the openneuro repository: https://openneuro.org/datasets/ds004498/versions/1.0.0. Data from the UKB dataset used here are available at https://www.fmrib.ox.ac.uk/ukbiobank/. The ABCD data used in this report came from ABCD BIDS Community Collection (ABCC; NDA Collection 3165) and the Annual Release 2.0: https://doi.org/10.15154/1503209. Data from the HCP dataset used here is available at www.humanconnectome.org. Users must agree to data use terms for the HCP before being allowed access to the data and ConnectomeDB, details are provided at https://www.humanconnectome.org/study/hcp-young-adult/data-use-terms. Data from the WU120 dataset is available in the openneuro repository at https://openneuro.org/datasets/ds000243/versions/00001. Data from the PRIME-DE Oxford macaque dataset used in this report are available at https://fcon_1000.projects.nitrc.org/indi/PRIME/oxford.html. Users register with NITRC and with the 1000 Functional Connectomes Project website on NITRC to gain access to the PRIME-DE datasets. Data from the UMN macaque will be publicly available via the PRIME-DE website (see above) by the end of 2023, after data collection of a larger sample is complete. Data from individual subjects participant 7 and 8, the individual neonate, infant and child participants, as well as those from the group average infant datasets, are available on reasonable request from C.J.L., E.M.G., J.R.P., C.M.S. and D.J.G. and C.D.S. They are not yet available through public databases because data collection is still ongoing. The DISTAL atlas is available from https://www.lead-dbs.org/helpsupport/knowledge-base/atlasesresources/distal-atlas/. The SUIT atlas is available from https://www.diedrichsenlab.org/imaging/suit.htm.

## Code availability

Task stimuli were presented using JSpsych toolbox v6.3, available from https://www.jspsych.org/6.3/. Data processing code for the ABCD data as well as for the child participant can be found at https://github.com/DCAN-Labs/abcd-hcp-pipeline. Data processing code for the HCP data can be found at https://github.com/Washington-University/HCPpipelines. Data processing code for participants 1–3 and 8, and the neonate, infant and perinatal stroke participants can be found at https://gitlab.com/DosenbachGreene/. Data processing code for participants 4–7 can be found at https://github.com/cjl2007/Liston-Laboratory-MultiEchofMRI-Pipeline. Data processing code for the macaque datasets can be found at https://github.com/DCAN-Labs/nhp-abcd-bids-pipeline. Code specific to the analyses in this manuscript can be found at https://gitlab.com/DosenbachGreene/SCAN/. Software packages incorporated into the above pipelines for data analysis included: Matlab R2020b, https://www.mathworks.com/; Connectome Workbench 1.5, http://www.humanconnectome.org/software/connectome-workbench.html; Freesurfer v6.2, https://surfer.nmr.mgh.harvard.edu/; FSL 6.0, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki; 4dfp tools, https://4dfp.readthedocs.io/en/latest/; and Infomap, https://www.mapequation.org.

# References


1. Penfield, W. & Boldrey, E. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain* **60**, 389–443 (1937).

2. Penfield, W. & Rasmussen, T. *The Cerebral Cortex of Man; A Clinical Study of Localization of Function*. xv, 248 (Macmillan, 1950).

3. Kwan, H. C., MacKay, W. A., Murphy, J. T. & Wong, Y. C. Spatial organization of precentral cortex in awake primates. II. Motor outputs. *J. Neurophysiol.* **41**, 1120–1131 (1978).

4. Graziano, M. S. A. Ethological action maps: a paradigm shift for the motor cortex. *Trends Cogn. Sci.* **20**, 121–132 (2016).

5. Dosenbach, N. U. F. et al. A core system for the implementation of task sets. *Neuron* **50**, 799–812 (2006).

6. Pool, J. L. & Ransohoff, J. Autonomic effects on stimulating rostral portion of cingulate gyri in man. *J. Neurophysiol.* **12**, 385–392 (1949).

7. Wall, P. D. & Davis, G. D. Three cerebral cortical systems affecting autonomic function. *J. Neurophysiol.* **14**, 507–517 (1951).

8. Neta, M. et al. Spatial and temporal characteristics of error-related activity in the human brain. *J. Neurosci.* **35**, 253–266 (2015).

9. Wager, T. D. et al. An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).

10. Dum, R. P., Levinthal, D. J. & Strick, P. L. Motor, cognitive, and affective areas of the cerebral cortex influence the adrenal medulla. *Proc. Natl Acad. Sci. USA* **113**, 9922–9927 (2016).

11. Desmurget, M. & Sirigu, A. Revealing humans' sensorimotor functions with electrical cortical stimulation. *Phil. Trans. R. Soc. B* **370**, 20140207 (2015).

12. Rathelot, J.-A. & Strick, P. L. Subdivisions of primary motor cortex based on cortico-motoneuronal cells. *Proc. Natl Acad. Sci. USA* **106**, 918–923 (2009).

13. Kuypers, H. G. J. M. & Brinkman, J. Precentral projections to different parts of the spinal intermediate zone in the rhesus monkey. *Brain Res.* **24**, 29–48 (1970).

14. Woolsey, C. N. et al. Patterns of localization in precentral and 'supplementary' motor areas and their relation to the concept of a premotor area. *Res. Publ. Assoc. Res. Nerv. Ment. Dis.* **30**, 238–264 (1952).

15. Picard, N. & Strick, P. L. Imaging the premotor areas. *Curr. Opin. Neurobiol.* **11**, 663–672 (2001).

16. Dum, R. P. & Strick, P. L. The origin of corticospinal projections from the premotor areas in the frontal lobe. *J. Neurosci.* **11**, 667–689 (1991).

17. Strick, P. L., Dum, R. P. & Rathelot, J.-A. The cortical motor areas and the emergence of motor skills: a neuroanatomical perspective. *Annu. Rev. Neurosci.* **44**, 425–447 (2021).

18. Bortoff, G. A. & Strick, P. L. Corticospinal terminations in two new-world primates: further evidence that corticomotoneuronal connections provide part of the neural substrate for manual dexterity. *J. Neurosci.* **13**, 5105–5118 (1993).

19. Rothwell, J. C. et al. Manual motor performance in a deafferented man. *Brain* **105**, 515–542 (1982).

20. Shadmehr, R., Smith, M. A. & Krakauer, J. W. Error correction, sensory prediction, and adaptation in motor control. *Annu. Rev. Neurosci.* **33**, 89–108 (2010).

21. Hoover, J. E. & Strick, P. L. The organization of cerebellar and basal ganglia outputs to primary motor cortex as revealed by retrograde transneuronal transport of herpes simplex virus type 1. *J. Neurosci.* **19**, 1446–1463 (1999).

22. Mink, J. W. The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* **50**, 381–425 (1996).

23. Power, J. D. et al. Functional network organization of the human brain. *Neuron* **72**, 665–678 (2011).

24. Gordon, E. M. et al. Precision functional mapping of individual human brains. *Neuron* **95**, 791–807 (2017).

25. Laumann, T. O. et al. Functional system and areal organization of a highly sampled individual human brain. *Neuron* **87**, 657–670 (2015).

26. Greene, D. J. et al. Integrative and network-specific connectivity of the basal ganglia and thalamus defined in individuals. *Neuron* **105**, 742–

758 (2020).

27. Marek, S. et al. Spatial and temporal organization of the individual human cerebellum. *Neuron* **100**, 977–993 (2018).

28. Kuehn, E. et al. Body topography parcellates human sensory and motor cortex. *Cereb. Cortex* **27**, 3790–3805 (2017).

29. Newbold, D. J. et al. Plasticity and spontaneous activity pulses in disused human brain circuits. *Neuron* **107**, 580–589 (2020).

30. Newbold, D. J. et al. Cingulo-opercular control network and disused motor circuits joined in standby mode. *Proc. Natl Acad. Sci. USA* **118**, e2019128118 (2021).

31. Barch, D. M. et al. Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80**, 169–189 (2013).

32. Laumann, T. O. et al. Brain network reorganisation in an adolescent after bilateral perinatal strokes. *Lancet Neurol.* **20**, 255–256 (2021).

33. Barrett, L. F. & Simmons, W. K. Interoceptive predictions in the brain. *Nat. Rev. Neurosci.* **16**, 419–429 (2015).

34. Mitra, A., Snyder, A. Z., Hacker, C. D. & Raichle, M. E. Lag structure in resting-state fMRI. *J. Neurophysiol.* **111**, 2374–2391 (2014).

35. Mitra, A. et al. Human cortical–hippocampal dialogue in wake and slow-wave sleep. *Proc. Natl Acad. Sci. USA* **113**, E6868–E6876 (2016).

36. Ohara, S. et al. Movement-related change of electrocorticographic activity in human supplementary motor area proper. *Brain* **123**, 1203–1215 (2000).

37. Yeo, B. T. T. et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).

38. Krubitzer, L., Huffman, K. J., Disbrow, E. & Recanzone, G. Organization of area 3a in macaque monkeys: contributions to the cortical phenotype. *J. Comp. Neurol.* **471**, 97–111 (2004).

39. Wagstyl, K. et al. BigBrain 3D atlas of cortical layers: cortical and laminar thickness gradients diverge in sensory and motor cortices. *PLoS Biol.* **18**, e3000678 (2020).

40. Correia, J. M., Caballero-Gaudes, C., Guediche, S. & Carreiras, M. Phonatory and articulatory representations of speech production in cortical and subcortical fMRI responses. *Sci. Rep.* **10**, 4529 (2020).

41. Eichert, N., Papp, D., Mars, R. B. & Watkins, K. E. Mapping human laryngeal motor cortex during vocalization. *Cereb. Cortex* **30**, 6254–6269 (2020).

42. Levinthal, D. J. & Strick, P. L. Multiple areas of the cerebral cortex influence the stomach. *Proc. Natl Acad. Sci. USA* **117**, 13078–13083 (2020).

43. Huber, L. et al. Sub-millimeter fMRI reveals multiple topographical digit representations that form action maps in human motor cortex. *NeuroImage* **208**, 116463 (2020).

44. Jensen, M. A. et al. Homunculus interruptus: a motor association area in the depth of the central sulcus. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.20.517292 (2022).

45. Ramsay, S. C. et al. Regional cerebral blood flow during volitional expiration in man: a comparison with volitional inspiration. *J. Physiol.* **461**, 85–101 (1993).

46. Sterling, P. Allostasis: a model of predictive regulation. *Physiol. Behav.* **106**, 5–15 (2012).

47. Woolsey, C. N., Erickson, T. C. & Gilson, W. E. Localization in somatic sensory and motor areas of human cerebral cortex as

determined by direct recording of evoked potentials and electrical stimulation. *J. Neurosurg.* **51**, 476–506 (1979).

48. Branco, D. M. et al. Functional variability of the human cortical motor map: electrical stimulation findings in perirolandic epilepsy surgery. *J. Clin. Neurophysiol.* **20**, 17–25 (2003).

49. Chassagnon, S., Minotti, L., Kremer, S., Hoffmann, D. & Kahane, P. Somatosensory, motor, and reaching/grasping responses to direct electrical stimulation of the human cingulate motor areas. *J. Neurosurg.* **109**, 593–604 (2008).

50. Desmurget, M. et al. Movement intention after parietal cortex stimulation in humans. *Science* **324**, 811–813 (2009).

51. Roux, F.-E., Niare, M., Charni, S., Giussani, C. & Durand, J.-B. Functional architecture of the motor homunculus detected by electrostimulation. *J. Physiol.* **598**, 5487–5504 (2020).

52. Borgomaneri, S., Serio, G. & Battaglia, S. Please, don't do it! Fifteen years of progress of non-invasive brain stimulation in action inhibition. *Cortex* **132**, 404–422 (2020).

53. Willett, F. R. et al. Hand knob area of premotor cortex represents the whole body in a compositional way. *Cell* **181**, 396–409.e26 (2020).

54. Silva, A. B. et al. A neurosurgical functional dissection of the middle precentral gyrus during speech production. *J. Neurosci.* **42**, 8416–8426 (2022).

55. Willett, F. R. et al. A high-performance speech neuroprosthesis. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.21.524489 (2023).

56. Muellbacher, W. et al. Improving hand function in chronic stroke. *Arch. Neurol.* **59**, 1278–1282 (2002).

57. Siegel, J. S. et al. The circuitry of abulia: Insights from functional connectivity MRI. *NeuroImage Clin.* **6**, 320–326 (2014).

58. Ohbayashi, M., Picard, N. & Strick, P. L. Inactivation of the dorsal premotor area disrupts internally generated, but not visually guided, sequential movements. *J. Neurosci.* **36**, 1971–1976 (2016).

59. Leyton, A. S. F. & Sherrington, C. S. Observations on the excitable cortex of the chimpanzee, orang-utan, and gorilla. *Q. J. Exp. Physiol.* **11**, 135–222 (1917).

60. Sherrington, S. C. S. *The Integrative Action of the Nervous System* (A. Constable, 1906).

61. Heller, A. et al. Arm function after stroke: measurement and recovery over the first three months. *J. Neurol. Neurosurg. Psychiatry* **50**, 714–719 (1987).

62. Dauer, W. & Przedborski, S. Parkinson's disease: mechanisms and models. *Neuron* **39**, 889–909 (2003).

63. DeYoe, E. A. et al. Mapping striate and extrastriate visual areas in human cerebral cortex. *Proc. Natl Acad. Sci. USA* **93**, 2382–2386 (1996).

64. Livingstone, M. & Hubel, D. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* **240**, 740–749 (1988).

65. Spiridon, M., Fischl, B. & Kanwisher, N. Location and spatial profile of category-specific regions in human extrastriate cortex. *Human Brain Mapping* **27**, 77–89 (2006).

66. Yi, H. G., Leonard, M. K. & Chang, E. F. The encoding of speech sounds in the superior temporal gyrus. *Neuron* **102**, 1096–1110 (2019).

67. Llinás, R. R. *I of the Vortex: From Neurons to Self*. x, 302 (The MIT Press, 2001).

68. Darby, R. R., Joutsa, J., Burke, M. J. & Fox, M. D. Lesion network localization of free will. *Proc. Natl Acad. Sci. USA* **115**, 10792–10797

(2018).

69. Seider, N. A. et al. Accuracy and reliability of diffusion imaging models. *NeuroImage* **254**, 119138 (2022).

70. Lynch, C. J. et al. Rapid precision functional mapping of individuals using multi-echo fMRI. *Cell Rep.* **33**, 108540 (2020).

71. Feczko, E. et al. Adolescent Brain Cognitive Development (ABCD) Community MRI collection and utilities. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.09.451638 (2021).

72. Lagore, R. L. et al. An 8-dipole transceive and 24-loop receive array for non-human primate head imaging at 10.5 T. *NMR Biomed.* **34**, e4472 (2021).

73. Moeller, S. et al. Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magn. Reson. Med.* **63**, 1144–1153 (2010).

74. Setsompop, K. et al. Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty. *Magn. Reson. Med.* **67**, 1210–1224 (2012).

75. Uğurbil, K. et al. Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project. *Neuroimage* **80**, 80–104 (2013).

76. Montez, D. F. et al. Using synthetic MR images for distortion correction. Preprint at *bioRxiv* https://doi.org/10.1101/2021.03.13.435270 (2022).

77. Noonan, M. P. et al. A neural circuit covarying with social hierarchy in macaques. *PLoS Biol.* **12**, e1001940 (2014).

78. Milham, M. P. et al. An open resource for non-human primate imaging. *Neuron* **100**, 61–74.e2 (2018).

79. Miller, K. L. et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1536 (2016).

80. Van Essen, D. C., Glasser, M. F., Dierker, D. L., Harwell, J. & Coalson, T. Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. *Cereb. Cortex* **22**, 2241–2262 (2012).

81. Glasser, M. F. et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).

82. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).

83. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).

84. Glasser, M. F. et al. The Human Connectome Project's neuroimaging approach. *Nat. Neurosci.* **19**, 1175–1187 (2016).

85. Robinson, E. C. et al. MSM: a new flexible framework for multimodal surface matching. *Neuroimage* **100**, 414–426 (2014).

86. Smith, S. M. et al. Resting-state fMRI in the Human Connectome Project. *NeuroImage* **80**, 144–168 (2013).

87. Gordon, E. M. et al. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cereb. Cortex* **26**, 288–303 (2016).

88. Sylvester, C. M. et al. Network-specific selectivity of functional connections in the neonatal brain. *Cereb. Cortex* **33**, 2200–2214 (2023).

89. Gordon, E. M. et al. Default-mode network streams for coupling to language and control systems. *Proc. Natl Acad. Sci. USA* **117**, 17308–

17319 (2020).

90. Gordon, E. M. et al. Individualized functional subnetworks connect human striatum and frontal cortex. *Cereb. Cortex* **32**, 2868–2884 (2022).

91. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage* **53**, 1–15 (2010).

92. Diedrichsen, J. & Zotow, E. Surface-based display of volume-averaged cerebellar imaging data. *PLoS ONE* **10**, e0133402 (2015).

93. Diedrichsen, J., Balsters, J. H., Flavell, J., Cussans, E. & Ramnani, N. A probabilistic MR atlas of the human cerebellum. *NeuroImage* **46**, 39–46 (2009).

94. Ewert, S. et al. Toward defining deep brain stimulation targets in MNI space: a subcortical atlas based on multimodal MRI, histology and structural connectivity. *NeuroImage* **170**, 271–282 (2018).

95. Chakravarty, M. M., Bertrand, G., Hodge, C. P., Sadikot, A. F. & Collins, D. L. The creation of a brain atlas for image guided neurosurgery using serial histological data. *NeuroImage* **30**, 359–376 (2006).

96. Raut, R. V., Mitra, A., Snyder, A. Z. & Raichle, M. E. On time delay estimation and sampling error in resting-state fMRI. *Neuroimage* **194**, 211–227 (2019).

97. Glasser, M. F. & Van Essen, D. C. Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *J. Neurosci.* **31**, 11597–11616 (2011).

98. Woolrich, M. W., Ripley, B. D., Brady, M. & Smith, S. M. Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage* **14**, 1370–1386 (2001).

# Acknowledgements

The authors thank P. Strick for extensive comments and discussions that allowed critical conceptualization of results; M. Graziano for helpful comments and suggestions; and J. Sallet, R. Mars and M. Rushworth for sharing their data with the PRIMatE Data Exchange consortium. This work was supported by NIH grants NS110332 (D.J.N.), MH120989 (C.J.L.), MH100019 (N.A.S.), MH129493 (D.M.B.), MH113883 (C.E.R.), MH128177 (J.Z.), EB031765 (J.Z.), DA048742 (J.Z.), MH120194 (J.T.W.), NS123345 (B.P.K.), NS098482 (B.P.K.), MH121518 (S.M.), MH128696 (T.X.), NS124789 (S.A.N.), MH118370 (C.G.), MH118362 (J.R.P.), HD088125 (J.R.P.), HD055741 (J.R.P.), MH121462 (J.R.P.), MH116961 (J.R.P.), MH129426 (J.R.P.), HD103525, MH120194 (J.T.W.), MH122389 (C.M.S.), DA047851 (C.J.L.), MH118388 (C.J.L.), MH114976 (C.J.L.), MH129616 (T.O.L.), DA041148 (D.A.F.), DA04112 (D.A.F.), MH115357 (D.A.F.), MH096773 (D.A.F. and N.U.F.D.), MH122066 (E.M.G., D.A.F. and N.U.F.D.), MH121276 (E.M.G., D.A.F. and N.U.F.D.), MH124567 (E.M.G., D.A.F. and N.U.F.D.), NS129521 (E.M.G., D.A.F. and N.U.F.D.), and NS088590 (N.U.F.D.); by NSF grant CAREER BCS-2048066 (C.G.); by Center for Brain Research in Mood Disorders; by Eagles Autism Challenge; by the Dystonia Medical Research Foundation (S.A.N.); by the National Spasmodic Dysphonia Association (E.M.G. and S.A.N.); by the Taylor Family Foundation (C.M.S. and T.O.L.); by the Intellectual and Developmental Disabilities Research Center (D.J.G. and N.U.F.D.); by the Kiwanis Foundation (N.U.F.D.); by the Washington University Hope Center for Neurological Disorders (E.M.G., B.P.K. and N.U.F.D.); and by Mallinckrodt Institute of Radiology pilot funding (D.J.G., E.M.G. and N.U.F.D.).

# Author information

## Authors and Affiliations

1. Mallinckrodt Institute of Radiology, Washington University School of Medicine, St Louis, MO, USA

Evan M. Gordon, Aishwarya Rajesh, Carolina Badke D'Andrea, Deanna M. Barch, Christopher D. Smyser, Joshua S. Shimony, Scott Marek, Scott A. Norris, Steven E. Petersen, Marcus E. Raichle & Nico U. F. Dosenbach

2. Department of Neurology, Washington University School of Medicine, St Louis, MO, USA

Roselyne J. Chauvin, Andrew N. Van, Ashley Nielsen, Dillan J. Newbold, Nicole A. Seider, Samuel R. Krimmel, Kristen M. Scheidter, Julia Monk, Ryland L. Miller, Athanasia Metoki, David F. Montez, Annie Zheng, Sydney Kaplan, Christopher D. Smyser, Jon T. Willie, Benjamin P. Kay, Scott A. Norris, Steven E. Petersen, Marcus E. Raichle & Nico U. F. Dosenbach

3. Department of Biomedical Engineering, Washington University in St. Louis, St Louis, MO, USA

Andrew N. Van, Peter Brunner, Steven E. Petersen, Marcus E. Raichle & Nico U. F. Dosenbach

4. Department of Neurology, New York University Langone Medical Center, New York, NY, USA

Dillan J. Newbold

5. Department of Psychiatry, Weill Cornell Medicine, New York, NY, USA

Charles J. Lynch, Immanuel Elbau, Cynthia E. Rogers, Jon T. Willie, Jonathan D. Power & Conor Liston

6. Department of Psychiatry, Washington University School of Medicine, St Louis, MO, USA

Nicole A. Seider, Ryland L. Miller, Tomoyuki Nishino, Michael J. Myers, Carolina Badke D'Andrea, Deanna M. Barch, Kelly N. Botteron, John R. Pruett, Chad M. Sylvester & Timothy O. Laumann

7. Department of Pediatrics, University of Minnesota, Minneapolis, MN, USA

Thomas Madison & Damien A. Fair

8. Department of Cognitive Science, University of California San Diego, La Jolla, CA, USA

Carolina Badke D'Andrea, Damion V. Demeter, Matthew Feigelis & Deanna J. Greene

9. Center for the Developing Brain, Child Mind Institute, New York, NY, USA

Julian S. B. Ramirez & Ting Xu

10. Department of Psychological and Brain Sciences, Washington University in St. Louis, St Louis, MO, USA

Deanna M. Barch, Steven E. Petersen, Marcus E. Raichle & Nico U. F. Dosenbach

11. Department of Pediatrics, Washington University School of Medicine, St Louis, MO, USA

Christopher D. Smyser, Cynthia E. Rogers & Nico U. F. Dosenbach

12. Department of Neuroscience, University of Minnesota, Minneapolis, MN, USA

Jan Zimmermann

13. Department of Neurosurgery, Washington University School of Medicine, St Louis, MO, USA

Jon T. Willie, Peter Brunner & Jarod L. Roland

14. Department of Psychology, Florida State University, Tallahassee, FL, USA

Caterina Gratton

15. Department of Neuroscience, Washington University School of Medicine, St Louis, MO, USA

    Steven E. Petersen & Marcus E. Raichle

16. Masonic Institute for the Developing Brain, University of Minnesota, Minneapolis, MN, USA

    Damien A. Fair

17. Institute of Child Development, University of Minnesota, Minneapolis, MN, 55455, United States

    Damien A. Fair

18. Program in Occupational Therapy, Washington University in St. Louis, St Louis, MO, USA

    Nico U. F. Dosenbach

## Contributions

Conception: E.M.G. and N.U.F.D. Design: E.M.G., R.J.C., T.O.L. and N.U.F.D. Data acquisition, analysis and interpretation: E.M.G., R.J.C., A.N.V., A.R., A.N., D.J.N., C.J.L., N.A.S., S.R.K., K.M.S., J.M., R.L.M., A.M., D.F.M., A.Z., I.E., T.M., T.N., M.J.M., S.K., C.B.D., D.V.D., M.F., J.S.B.R., T.X., D.M.B., C.D.S., C.E.R., J.Z., K.N.B., J.R.P., J.T.W., P.B., J.S.S., B.P.K., S.M., S.A.N., C.G., C.M.S., J.D.P., C.L., D.J.G., J.L.R., S.E.P., M.E.R., T.O.L., D.A.F. and N.U.F.D. Manuscript writing and revision: E.M.G., D.J.N., D.M.B., S.M., S.A.N., C.G., C.M.S., J.D.P., D.J.G., S.E.P., M.E.R., T.O.L., D.A.F., and N.U.F.D. Participant 1 was author A.N; participant 2 was author N.U.F.D; participant 4 was author C.J.L; participant 5 was author J.D.P; and participant 8 was author E.M.G.

## Corresponding authors

Correspondence to Evan M. Gordon or Nico U. F. Dosenbach.

# Ethics declarations

## Competing interests

D.A.F., N.U.F.D. and N.A.S. have a financial interest in Turing Medical Inc. and may benefit financially if the company is successful in marketing FIRMM motion monitoring software products. A.N.V., D.A.F. and N.U.F.D. may receive royalty income based on FIRMM technology developed at Washington University School of Medicine and Oregon Health and Sciences University and licensed to Turing Medical Inc. D.A.F. and N.U.F.D. are co-founders of Turing Medical Inc. These potential conflicts of interest have been reviewed and are managed by Washington University School of Medicine, Oregon Health and Sciences University and the University of Minnesota. N.A.S. is now an employee of Turing Medical Inc. C.M.S receives research support from Sage Therapeutics. C.L. is listed as an inventor for Cornell University patent applications on neuroimaging biomarkers for depression that are pending or in preparation. C.L. has served as a scientific advisor or consultant to Compass Pathways PLC, Delix Therapeutics, Magnus Medical and Brainify.AI. The other authors declare no competing interests.

# Peer review

## Peer review information

*Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

# Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# Extended data figures and tables

## Extended Data Fig. 1 Consistency of the inter-effector motif across datasets.

Connectivity patterns seeded from a continuous line down the left precentral gyrus revealed that the interleaved motor functional connectivity pattern was consistent across **a**, seven highly-sampled individual participants (172–356 min of data); **b**, replication data (416–1,114 min) collected in P1–P3; and **c**, multiple independent sets of group data averaged across cohorts of varying size. Here, functional connectivity is shown seeded from the middle inter-effector region for each individual participant and group-averaged dataset (see Supplementary Video 2 for all seeds). Thresholds for connectivity maps were scaled to the 95$^{th}$ percentile of map values in individuals, and to the 97$^{th}$ percentile of values in groups, to account for differences in data acquisition and processing strategies across datasets. **d**, Discrete functional networks were demarcated within each subject in M1 and S1 using a whole-brain, data-driven hierarchical approach applied to the resting-state fMRI data (see Fig. S7), which defined the spatial extent of the networks observed in Fig. 1 (black outlines). In P1-P3, regions defined by resting state functional connectivity (RSFC) were functionally labeled using a classic block-design fMRI motor task involving separate movement of the foot, hand, and tongue (following[31]; see[29] for details). The map illustrates the top 1% of vertices activated by movement of the foot (green), hand (cyan), and mouth (orange). **e**, Left: preferential connectivity of each motor division to the cerebellum. Right: activations during the fMRI motor task described in panel d. The map illustrates the top 5% of vertices within cerebellum active during movement of the foot (green), hand (cyan), and mouth (orange).

## Extended Data Fig. 2 Motor cortex functional connectivity in pediatric participants and perinatal stroke.

Functional connectivity maps were seeded from a continuous line of points down precentral gyrus in fMRI data from **a**, data averaged across 262 human neonates, all scanned shortly after birth; **b**, a neonate scanned 13

days after birth; **c**, an 11-month old infant; **d**, a 9-year old child; **e**, adult participant P1 (from Fig. 1); and **f**, an adolescent who had experienced extensive cortical reorganization after severe bilateral perinatal strokes (destroyed cortex in black). Right hemisphere is shown in the stroke patient because left hemisphere M1 was entirely lost. Example seed maps shown here illustrate observed inter-effector (row 1) and effector-specific connectivity (rows 2-4). Inter-effector and effector-specific regions exhibited clear boundaries within M1 in the infant, child, the adults, and the stroke patient, but not in the neonates. Visualization thresholds varied between $Z(r) > 0.3$ and $Z(r) > 0.5$ across datasets due to differences in data collection and processing, as well as differences inherent to the populations.

## Extended Data Fig. 3 Whole brain functional connectivity of inter-effector motif across participants.

Brain regions with the strongest functional connectivity to the middle inter-effector region in **a**, medial cortex, **b**, striatum (lateral view of left and right striatum), **c**, thalamus (axial view), and **d**, cerebellum. Functional connectivity values are thresholded at $Z(r) > 0.35$ in cortex. Subcortical functional connectivity values are thresholded at different levels in each subject due to variation in subcortical signal-to-noise ratios across individuals. Thresholds were chosen to illustrate the strongest subcortical connections. Specific thresholds shown here: P1 - $Z(r) > 0.15$; P3, 4, 6, 7 - $Z(r) > 0.1$; P2 - $Z(r) > 0.04$; P5 - $Z(r) > 0.03$.

## Extended Data Fig. 4 Functional connectivity and structural MRI metrics of motor cortex regions.

In each individual participant, measures derived from each of the foot, hand, mouth, and inter-effector motor regions. Colored lines connect the same participant's inter-effector and effector-specific regions for ease of comparison. **a**, Functional connectivity strength $Z(r)$ between M1 region and individual-specific Cingulo-Opercular Network (CON). **b**, Functional connectivity between M1 region and middle insula. **c**, Functional connectivity with Lobule VIIIa vermis of the cerebellum. **d**, Functional connectivity between M1 region and dorsal posterior putamen. **e—g**,

Functional connectivity between M1 region and nuclei of the thalamus: **e**, Centromedian nucleus; **f**, Ventral Intermediate nucleus; **g**, Ventral Posteromedial nucleus. **h**, Functional connectivity between M1 region and adjacent postcentral gyrus (S1). **i**, Cortical thickness in M1 region. **j**, Fractional Anisotropy within 2 mm below cortex under M1 region. **k**, Intracortical myelin, indexed by the T1/T2 ratio and normalized across cortex, within cortex of M1 region. All significance values reflect significance across three two-sided paired t-tests (inter-effector vs foot, vs hand, and vs mouth). * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$, FDR-corrected.

## Extended Data Fig. 5 Differences in functional connectivity between inter-effector regions.

Brain regions more strongly connected to the superior inter-effector region than to either of the other two (top row); relatively most strongly connected to the middle inter-effector region (middle row); and relatively most strongly connected to the inferior inter-effector region, in cortex (left), striatum, thalamus, and cerebellum (right), **a**, in at least 50% of individuals ($n = 7$) and **b**, in group-averaged data from the Human Connectome Project (HCP; $n = 812$). Thresholds used are the same as in Fig. 2b. Note that central sulcus regions are masked as they exhibit large differences by definition. See Fig. S3 for all individual participants.

## Extended Data Fig. 6 Effector-specific and inter-effector and regions in pre- and postcentral gyrus.

In every participant, Brodmann Areas (BAs) in M1 (BAs 4a, 4p) and S1 (BAs 1, 2, 3a, 3b) are displayed on the cerebral cortex, tilted around the Y- and Z-axes to show S1. Overlaid are **a**, the somatomotor-hand region, and **b**, the inter-effector regions.

## Extended Data Fig. 7 Primary motor cortex activation profiles for movement task battery.

In two participants (top, bottom), LOWESS curves were fit to the task activation profiles at each dorsal-ventral point in M1, for each separate movement (colored lines). Colored blocks (top) show the effector-specific foot (green), hand (cyan), and mouth (orange) areas of M1, as well as the inter-effector regions (maroon); dotted maroon lines show the boundaries between regions. The centers of effector-specific regions are characterized by strong activations for movements of the most distal body parts, and deactivations for all other movements. Inter-effector regions, by contrast, exhibited moderate activations for most movements.

## Extended Data Fig. 8 Effector-specificity of task fMRI activations.

In each participant, in the **a**, abdominal flexure task and the **b**, eyebrow raising task, the inter-effector regions and cingulo-opercular network (CON) were active. By contrast, in **c**, toe and **d**, hand motion tasks, activation was much more specific to a single region of somatomotor cortex. **e**, Across tasks, the degree of CON activation was consistently similar to the activation of the inter-effector regions (correlation between CON and inter-effector activations: all Pearson's $r > 0.81$, $P < 10^{-5}$, FDR corrected), but not consistently to hand (CON vs hand: Pearson's $r > 0.05$, $P < 0.82$) or foot (CON vs foot: Pearson's $r > 0.33$, $P < 0.13$) regions, and more weakly to mouth regions (CON vs mouth: Pearson's $r > 0.61$, $P < 0.003$). Illustrated activation values are averaged across participants and ordered based on CON activation.

## Extended Data Fig. 9 Motor Cortex functional connectivity in non-human primates.

Functional connectivity maps were seeded from dorsal anterior cingulate cortex (top row), as well as from a continuous line of points down anterior central sulcus (rows 2-4), in fMRI data from **a**, an individual macaque scanned for 77 min on a 10.5T MRI scanner; **b**, an individual macaque scanned for 53 min on a 3T scanner; and **c**, group-averaged data from eight macaques each scanned for 53 min on a 3T scanner. The dorsal anterior cingulate seed demonstrated connectivity to frontal, insular, and parietal

regions homologous with the human CON, as well as with two regions in anterior central sulcus (maroon arrows). These central sulcus regions are thought to correspond to areas that project to internal organs[10] and represent possible macaque homologues of the inter-effector regions. The central sulcus seeds demonstrated connectivity patterns corresponding to the known functional divisions between M1 regions representing the foot (second row), hand (third row), and face (bottom row).

### Extended Data Fig. 10 Somato-Cognitive Action Network regions in human cortical surface stimulation data.

The map of somato-cognitive action network (SCAN) regions was compared with published movements evoked by direct cortical surface stimulation[51]. Cortical map: functional connectivity is shown seeded from the middle SCAN region and averaged across all subjects in the HCP dataset ($n = 812$; see also Extended Data Fig. 1c). Stimulation locations: MNI coordinates of surface stimulation location, and the resulting evoked movement, from 100 patients undergoing awake surgical brain mapping were reported in[51]. Each stimulation location evoking movement was mapped to the nearest cortical vertex on a group-averaged pial surface. Stimulation sites are colored according to whether they evoked facial movements (orange) or upper extremity movements (cyan). Stimulation sites evoking movement did not overlap with the central inter-effector region.

## Supplementary information

### Supplementary Information

This file contains Supplementary Tables 1 and 2 and Supplementary Figs. 1–8.

### Reporting Summary

### Peer Review File

[Supplementary Video 1]

[Supplementary Video 2]

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

## About this article

### Cite this article

Gordon, E.M., Chauvin, R.J., Van, A.N. *et al.* A somato-cognitive action network alternates with effector regions in motor cortex. *Nature* **617**, 351–359 (2023). https://doi.org/10.1038/s41586-023-05964-2

- Received: 02 November 2022
- Accepted: 16 March 2023
- Published: 19 April 2023
- Issue Date: 11 May 2023

- DOI: https://doi.org/10.1038/s41586-023-05964-2

### Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

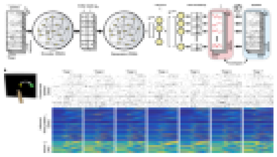
## A redrawn map for the human motor cortex

- David A. Leopold

Nature News & Views 19 Apr 2023

---

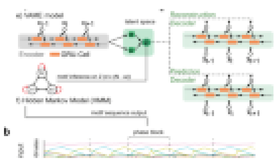
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05964-2



## Inferring single-trial neural population dynamics using sequential auto-encoders

17 September 2018

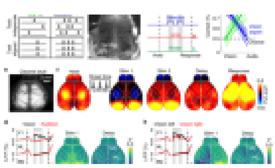
Chethan Pandarinath, Daniel J. O'Shea, … David Sussillo



## Identifying behavioral structure from deep variational embeddings of animal motion

18 November 2022

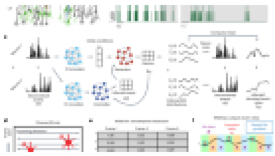
Kevin Luxem, Petra Mocellin, … Pavol Bauer



## Single-trial neural dynamics are dominated by richly varied movements

24 September 2019

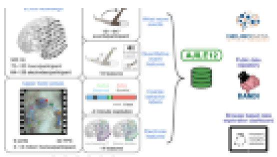
Simon Musall, Matthew T. Kaufman, … Anne K. Churchland



## [A deep learning framework for inference of single-trial neural population dynamics from calcium imaging with subframe temporal resolution]

24 November 2022

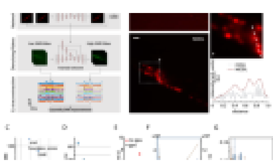
Feng Zhu, Harrison A. Grier, … Chethan Pandarinath



## [AJILE12: Long-term naturalistic human intracranial neural recordings and pose]

21 April 2022

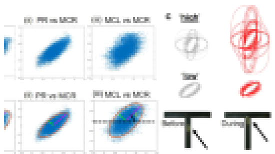
Steven M. Peterson, Satpreet H. Singh, … Bingni W. Brunton



## [Fast, efficient, and accurate neuro-imaging denoising via supervised deep learning]

02 September 2022

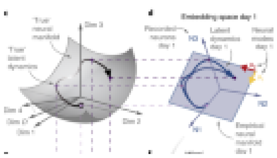
Shivesh Chaudhary, Sihoon Moon & Hang Lu



## [High resolution behavioral and neural activity representation using a geometrical approach]

14 May 2020

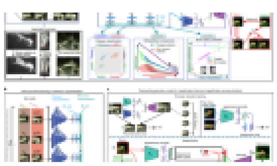
Zev Brand & Avi Avital



## [Long-term stability of cortical population dynamics underlying consistent behavior]

06 January 2020

Juan A. Gallego, Matthew G. Perich, … Lee E. Miller



## [Unsupervised behaviour analysis and magnification (uBAM) using deep learning]

05 April 2021

Biagio Brattoli, Uta Büchler, … Björn Ommer

- Article
- Open Access
- [Published: 03 May 2023]

# Learnable latent embeddings for joint behavioural and neural analysis


- Steffen Schneider,
- Jin Hwa Lee &
- Mackenzie Weygandt Mathis


*Nature* **volume 617**, pages 360–368 (2023)


- 54k Accesses

- 1004 Altmetric

- Metrics details


## Abstract

Mapping behavioural actions to neural activity is a fundamental goal of neuroscience. As our ability to record large neural and behavioural data increases, there is growing interest in modelling neural dynamics during adaptive behaviours to probe neural representations[1,2,3]. In particular, although neural latent embeddings can reveal underlying correlates of behaviour, we lack nonlinear techniques that can explicitly and flexibly leverage joint behaviour and neural data to uncover neural dynamics[3,4,5]. Here, we fill this gap with a new encoding method, CEBRA, that jointly uses behavioural and neural data in a (supervised) hypothesis- or (self-supervised) discovery-driven manner to produce both consistent and high-performance latent spaces. We show that consistency can be used as a metric for uncovering meaningful differences, and the inferred latents can be used for decoding. We validate its accuracy and demonstrate our tool's utility for both calcium and electrophysiology datasets, across sensory and motor tasks and in simple or complex behaviours across species. It allows leverage of single- and multi-session datasets for hypothesis testing or can be used label

free. Lastly, we show that CEBRA can be used for the mapping of space, uncovering complex kinematic features, for the production of consistent latent spaces across two-photon and Neuropixels data, and can provide rapid, high-accuracy decoding of natural videos from visual cortex.

# Main

A central quest in neuroscience is the neural origin of behaviour[1,2]. Nevertheless, we are still limited in both the number of neurons and length of time we can record from behaving animals in a session. Therefore, we need new methods that can combine data across animals and sessions with minimal assumptions, thereby generating interpretable neural embedding spaces[1,3]. Current tools for representation learning are either linear or, if nonlinear, typically rely on generative models and they do not yield consistent embeddings across animals (or repeated runs of the algorithm). Here, we combine recent advances in nonlinear disentangled representation learning and self-supervised learning to develop a new dimensionality reduction method that can be applied jointly to behavioural and neural recordings to show meaningful lower-dimensional neural population dynamics[3,4,5].

From data visualization (clustering) to discovery of latent spaces that explain neural variance, dimensionality reduction of behaviour or neural data has been impactful in neuroscience. For example, complex three-dimensional (3D) forelimb reaching can be reduced to between only eight and twelve dimensions[6,7], and low-dimensional embeddings show some robust aspects of movements (for example, principal component analysis (PCA)-based manifolds in which the neural state space can easily be constrained and is stable across time[8,9,10]). Linear methods such as PCA are often used to increase interpretability, but this comes at the cost of performance[1]. Uniform manifold approximation and projection (UMAP)[11] and *t*-distributed stochastic neighbour embedding (*t*-SNE)[12] are excellent nonlinear methods but they lack the ability to explicitly use time information, which is always available in neural recordings, and they are not as directly interpretable as PCA. Nonlinear methods are desirable for use in high-performance decoding but often lack identifiability—the desirable property that true model

parameters can be determined, up to a known indeterminacy[13,14]. This is critical because it ensures that the learned representations are uniquely determined and thus facilitates consistency across animals and/or sessions.

There is recent evidence that label-guided variational auto-encoders (VAEs) could improve interpretability[5,15,16]. Namely, by using behavioural variables, such algorithms can learn to project future behaviour onto past neural activity[15], or explicitly to use label priors to shape the embedding[5]. However, these methods still have restrictive explicit assumptions on the underlying statistics of the data and they do not guarantee consistent neural embeddings across animals[5,17,18], which limits both their generalizability and interpretability (and thereby affects accurate decoding across animals).

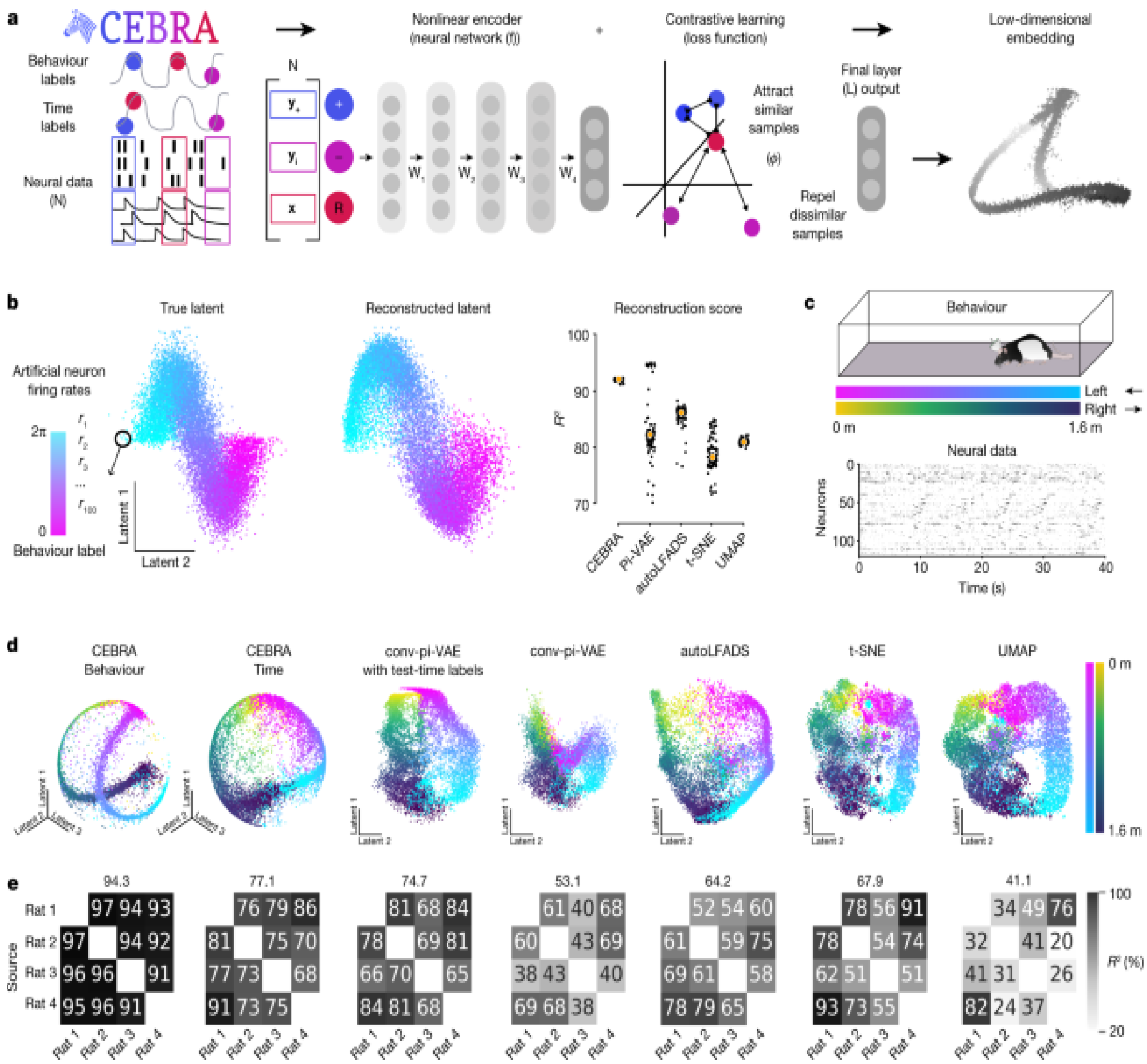
We address these open challenges with CEBRA, a new self-supervised learning algorithm for obtaining interpretable, consistent embeddings of high-dimensional recordings using auxiliary variables. Our method combines ideas from nonlinear independent component analysis (ICA) with contrastive learning[14,19,20,21], a powerful self-supervised learning scheme, to generate latent embeddings conditioned on behaviour (auxiliary variables) and/or time. CEBRA uses a new data-sampling scheme to train a neural network encoder with a contrastive optimization objective to shape the embedding space. It can also generate embeddings across multiple subjects and cope with distribution shifts among experimental sessions, subjects and recording modalities. Importantly, our method relies on neither data augmentation (as does SimCLR[22]) nor a specific generative model, which would limit its range of use.

# Joint behavioural and neural embeddings

We propose a framework for jointly trained latent embeddings. CEBRA leverages user-defined labels (supervised, hypothesis-driven) or time-only labels (self-supervised, discovery-driven; Fig. 1a and Supplementary Note 1) to obtain consistent embeddings of neural activity that can be used for both visualization of data and downstream tasks such as decoding. Specifically, it is an instantiation of nonlinear ICA based on contrastive learning[14]. Contrastive learning is a technique that leverages contrasting samples (positive and negative) against each other to find attributes in common and

those that separate them. We can use discrete and continuous variables and/or time to shape the distribution of positive and negative pairs, and then use a nonlinear encoder (here, a convolutional neural network but can be another type of model) trained with a new contrastive learning objective. The encoder features form a low-dimensional embedding of the data (Fig. 1a). Generation of consistent embeddings is highly desirable and closely linked to identifiability in nonlinear ICA[14,23]. Theoretical work has shown that the use of contrastive learning with auxiliary variables is identifiable for bijective neural networks using a noise contrastive estimation (NCE) loss[14], and that with an InfoNCE loss this bijectivity assumption can sometimes be removed[24] (see also our theoretical generalization in Supplementary Note 2). InfoNCE minimization can be viewed as a classification problem such that, given a reference sample, the correct positive sample needs to be distinguished from multiple negative samples.

**Fig. 1: Use of CEBRA for consistent and interpretable embeddings.**




**a**, CEBRA allows for self-supervised, supervised and hybrid approaches for both hypothesis-and discovery-driven analysis. Overview of pipeline: collect data (for example, pairs of behaviour (or time) and neural data (**x**,**y**)), determine positive and negative pairs, train CEBRA and produce embeddings. $W_{1,...4}$ represent the neural network weights. **b**, Left, true 2D latent, where each point is mapped to the spiking rate of 100 neurons. Middle, CEBRA embedding after linear regression to the true latent. Right, reconstruction score is $R^2$ of linear regression between the true latent and resulting embedding from each method. The ‘behaviour label’ is a 1D random variable sampled from uniform distribution of [0, 2π] that is assigned to each time bin of synthetic neural data, as visualized by the colour map. The orange line represents the median and each black dot an individual run ($n$ = 100). CEBRA-Behaviour shows a significantly higher

reconstruction score compared with pi-VAE, *t*-SNE and UMAP (one-way ANOVA, $F(4, 495) = 251$, $P = 1.12 \times 10^{-117}$ with post hoc Tukey's honest significant difference $P < 0.001$). **c**, Rat hippocampus data derived from ref. [26]. Cartoon from scidraw.io. Electrophysiology data were collected while a rat traversed a 1.6 m linear track 'leftwards' or 'rightwards'. **d**, We benchmarked CEBRA against conv-pi-VAE (both with labels and without), autoLFADS, *t*-SNE and unsupervised UMAP. Note: for performance against the original pi-VAE see Extended Data Fig. 1. We plot the three latents (all CEBRA-embedding figures show the first three latents). The dimensionality of the latent space is set to the minimum and equivalent dimension per method (3D for CEBRA and 2D for others) for fair comparison. Note: higher dimensions for CEBRA can yield higher consistency values (Extended Data Fig. 7). **e**, Correlation matrices show $R^2$ values after fitting a linear model between behaviour-aligned embeddings of pairs of rats, one as the target and the other as the source (mean, $n = 10$ runs). Parameters were picked by optimization of average run consistency across rats.

CEBRA optimizes neural networks **f**, **f′** that map neural activity to an embedding space of a defined dimension (Fig. 1a). Pairs of data (**x**, **y**) are mapped to this embedding space and then compared with a similarity measure $\phi(\cdot,\cdot)$. Abbreviating this process with $\psi \left({\bf{x}},{\bf{y}}\right)=\varphi \left({\bf{f}}\left({\bf{x}}\,\right),{{\bf{f}}}^{{\prime} }\left({\bf{y}}\right)\right)/\tau$ and a temperature hyperparameter, $\tau$, the full criterion for optimization is

$$\mathop{{\mathbb{E}}}\limits_{\begin{array}{c}{\bf{x}}\sim p({\bf{x}}),\,{{\bf{y}}}_{+}\sim p({\bf{y}}|{\bf{x}})\\ {{\bf{y}}}_{1},\ldots ,{{\bf{y}}}_{n}\sim q({\bf{y}}|{\bf{x}})\end{array}}\,[\,-\psi ({\bf{x}},{{\bf{y}}}_{+})+\log \mathop{\sum }\limits_{i=1}^{n}{e}^{\psi ({\bf{x}},{{\bf{y}}}_{i})}],$$

which, depending on the dataset size, can be optimized with algorithms for either batch or stochastic gradient descent.

In contrast to other contrastive learning algorithms, the positive-pair distribution $p$ and negative-pair distribution $q$ can be systematically designed and allow the use of time, behaviour and other auxiliary information to shape the geometry of the embedding space. If only discrete labels are used,

this training scheme is conceptually similar to supervised contrastive learning[21].

CEBRA can leverage continuous behavioural (kinematics, actions) as well as other discrete variables (trial ID, rewards, brain-area ID and so on). Additionally, user-defined information about desired invariances in the embedding is used (across animals, sessions and so on), allowing for flexibility in data analysis. We group this information into task-irrelevant and -relevant variables, and these can be leveraged in different contexts. For example, to investigate trial-to-trial variability or learning across trials, information such as a trial ID would be considered a task-relevant variable. On the contrary, if we aim to build a robust brain machine interface that should be invariant to such short-term changes, we would include trial information as a task-irrelevant variable and obtain an embedding space that no longer carries this information. Crucially, this allows inference of latent embeddings without explicit modelling of the data-generating process (as done in pi-VAE[5] and latent factor analysis via dynamical systems (LFADS)[17]). Omitting the generative model and replacing it by a contrastive learning algorithm facilitates broader applicability without modifications.

# Robust and decodable latent embeddings

We first demonstrate that CEBRA significantly outperforms *t*-SNE, UMAP, automatic LFADS (autoLFADS)[25] and pi-VAE (the latter was shown to outperform PCA, LFADS, demixed PCA and PfLDS (Poisson feed-forward neural network linear dynamical system) on some tasks) in the reconstruction of ground truth synthetic data (one-way analysis of variance (ANOVA), $F(4, 495) = 251$, $P = 1.12 \times 10^{-117}$; Fig. 1b and Extended Data Fig. 1a,b).

We then turned to a hippocampus dataset that was used to benchmark neural embedding algorithms[5,26] (Extended Data Fig. 1c and Supplementary Note 1). Of note, we first significantly improved pi-VAE by the addition of a convolutional neural network (conv-pi-VAE), thereby allowing this model to leverage multiple time steps, and used this for further benchmarking (Extended Data Fig. 1d,e). To test our methods, we first considered the correlation of the resulting embedding space across subjects (does it produce

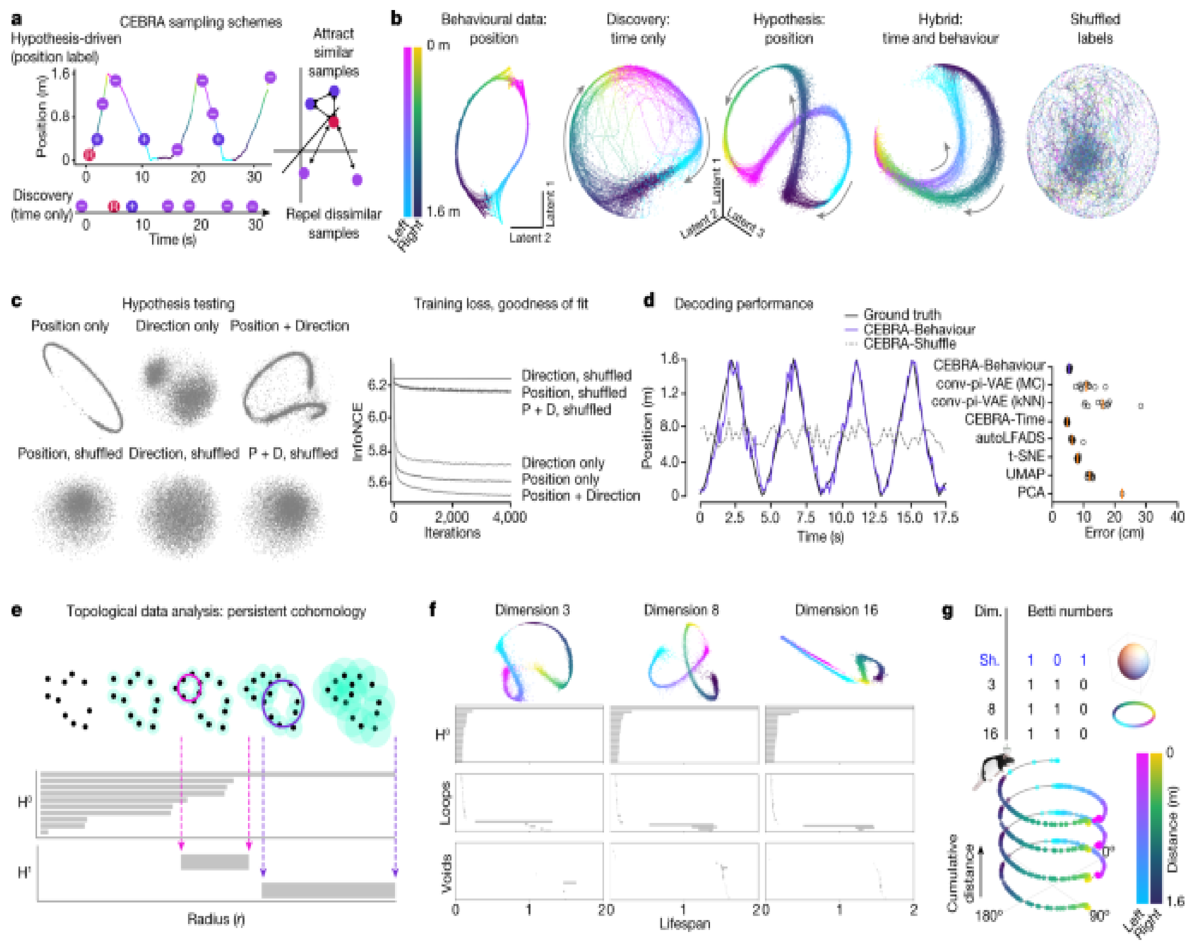
similar latent spaces?), and the correlation across repeated runs of the algorithm (how consistent are the results?). We found that CEBRA significantly outperformed other algorithms in the production of consistent embeddings, and it produced visually informative embeddings (Fig. 1c–e and Extended Data Figs. 2 and 3; for each embedding a single point represents the neural population activity over a specified time bin).

When using CEBRA-Behaviour, the consistency of the resulting embedding space across subjects is significantly higher compared with autoLFADS and conv-pi-VAE, with or without test-time labels (one-way ANOVA $F(25.4)$ $P = 1.92 \times 10^{-16}$; Supplementary Table 1 and Fig. 1d,e). Qualitatively, it can be appreciated that both CEBRA-Behaviour and -Time have similar output embeddings whereas the latents from conv-pi-VAE, either with label priors or without labels, are not consistent (CEBRA does not need test-time labels), suggesting that the label prior strongly shapes the output embedding structure of conv-pi-VAE. We also considered correlations across repeated runs of the algorithm, and found higher consistency and lower variability with CEBRA (Extended Data Fig. 4).

# Hypothesis-driven and discovery-driven analyses

Among the advantages of CEBRA are its collective flexibility, limited assumptions, and ability to test hypotheses. For the hippocampus, one can hypothesize that these neurons represent space[27,28] and therefore the behavioural label could be either position or velocity (Fig. 2a). In addition, considering structure in only the behavioural data (with CEBRA) could help refine which behavioural labels to use jointly with neural data (Fig. 2b). Conversely, for the sake of argument, we could have an alternative hypothesis: that the hippocampus does not map space, but simply maps the direction of travel or some other feature. Using the same model but hypothesis free, and using time for selection of contrastive pairs, is also possible, and/or a hybrid thereof (Fig. 2a,b).

**Fig. 2: Hypothesis- and discovery-driven analysis with CEBRA.**




**a**, CEBRA can be used in any of three modes: hypothesis-driven mode, discovery-driven mode, or hybrid mode, which allows for weaker priors on the latent embedding. **b**, Left to right, CEBRA on behavioural data using position as a label, CEBRA-Time, CEBRA-Behaviour (on neural data) with position hypothesis, CEBRA-Hybrid (a five-dimensional space was used, in which 3D is first guided by both behaviour + time and the final 2D is guided by time) and shuffled (erroneous). **c**, Embeddings with position (P) only, direction (D) only and P + D only, and shuffled position only, direction only and P + D only, for hypothesis testing. The loss function can be used as a metric for embedding quality. **d**, Left, we utilized either hypothesis-driven P + D or shuffle (erroneous) to decode the position of the rat, which yielded a large difference in decoding performance: P + D $R^2$ = 73.35 versus −49.90% for shuffled, and median absolute error 5.8 versus 44.7 cm. Purple line represents decoding from the 3D hypothesis-based latent space; dashed line is shuffled. Right, performance across additional methods (orange bars indicate the median of individual runs ($n$ = 10), indicated by black circles.

Each run is averaged over three splits of the dataset). MC, Monte Carlo. **e**, Schematic showing how persistent cohomology is computed. Each data point is thickened to a ball of gradually expanding radius (*r*) while tracking the birth and death of 'cycles' in each dimension. Prominent lifespans, indicated by pink and purple arrows, are considered to determine Betti numbers. **f**, Top, visualization of neural embeddings computed with different input dimensions. Bottom, related persistent cohomology lifespan diagrams. **g**, Betti numbers from shuffled embeddings (sh.) and across increasing dimensions (dim.) of CEBRA, and the topology-preserving circular coordinates using the first cocycle from persistent cohomology analysis (Methods).

We trained hypothesis-guided (supervised), time-only (self-supervised) and hybrid models across a range of input dimensions and embedded the neural latents into a 3D space for visualization. Qualitatively, we find that the position-based model produces a highly smooth embedding that shows the position of the animal—namely, there is a continuous 'loop' of latent dynamics around the track (Fig. 2b). This is consistent with what is known about the hippocampus[26] and shows the topology of the linear track with direction specificity whereas shuffling the labels, which breaks the correlation between neural activity and direction and position, produces an unstructured embedding (Fig. 2b).

CEBRA-Time produces an embedding that more closely resembles that of position (Fig. 2b). This also suggests that time contrastive learning captured the major latent space structure, independent of any label input, reinforcing the idea that CEBRA can serve both discovery- and hypothesis-driven questions (and that running both variants can be informative). The hybrid design, whose goal is to disentangle the latent to subspaces that are relevant to the given behavioural and residual temporal variance and noise, showed a structured embedding space similar to behaviour (Fig. 2b).

To quantify how CEBRA can disentangle which variable had the largest influence on embedding, we tested for encoding position, direction and combinations thereof (Fig. 2c). We find that position plus direction is the most informative label[29] (Fig. 2c and Extended Data Fig. 5a–d). This is evident both in the embedding and the value of the loss function on convergence, which serves as a 'goodness of fit' metric to select the best

labels—that is, which label(s) produce the lowest loss at the same point in training (Extended Data Fig. 5e). Note that erroneous (shuffled) labels converge to considerably higher loss values.

To measure performance, we consider how well we could decode behaviour from the embeddings. As an additional baseline we performed linear dimensionality reduction with PCA. We used a *k*-nearest-neighbour (kNN) decoder for position and direction and measured the reconstruction error. We find that CEBRA-Behaviour has significantly better decoding performance (Fig. 2d and Supplementary Video 1) compared with both pi-VAE and our conv-pi-VAE (one-way ANOVA, $F = 131$, $P = 3.6 \times 10^{-24}$), and also CEBRA-Time compared with unsupervised methods (autoLFADS, *t*-SNE, UMAP and PCA; one-way ANOVA, $F = 1{,}983$, $P = 6 \times 10^{-50}$; Supplementary Table 2). Zhou and Wei[5] reported a median absolute decoding error of 12 cm error whereas we achieved approximately 5 cm (Fig. 2d). CEBRA therefore allows for high-performance decoding and also ensures consistent embeddings.

# Cohomology as a metric for robustness

Although CEBRA can be trained across a range of dimensions, and models can be selected based on decoding, goodness of fit and consistency, we also sought to find a principled approach to verify the robustness of embeddings that might yield insight into neural computations[30,31] (Fig. 2e). We used algebraic topology to measure the persistent cohomology as a comparison in regard to whether learned latent spaces are equivalent. Although it is not required to project embeddings onto a sphere, this has the advantage that there are default Betti numbers (for a *d*-dimensional uniform embedding, \({H}^{0}=1,{H}^{1}=0,\cdots ,{H}^{d-1}=1\)—that is, 1,0,1 for the two-sphere). We used the distance from the unity line (and threshold based on a computed null shuffled distribution in Births versus Deaths to compute Betti numbers; Extended Data Fig. 6). Using CEBRA-Behaviour or -Time we find a ring topology (1,1,0; Fig. 2f), as one would expect from a linear track for place cells. We then computed the Eilenberg–MacLane coordinates for the identified cocycle ($H^1$) for each model[32,33]—this allowed us to map each time point to topology-preserving coordinates—and indeed we find that the ring topology for the CEBRA models matches space (position) across

dimensions (Fig. 2g and Extended Data Fig. 6). Note that this topology differs from (1,0,1)—that is, Betti numbers for a uniformly covered sphere —which in our setting would indicate a random embedding as found by shuffling (Fig. 2g).

# Multi-session, multi-animal CEBRA

CEBRA can also be used to jointly train across sessions and different animals, which can be highly advantageous when there is limited access to simultaneously recorded neurons or when looking for animal-invariant features in the neural data. We trained CEBRA across animals within each multi-animal dataset and find that this joint embedding allows for even more consistent embeddings across subjects (Extended Data Fig. 7a–c; one-sided, paired $t$-tests; Allen data: $t = -5.80$, $P = 5.99 \times 10^{-5}$; hippocampus: $t = -2.22$, $P = 0.024$).

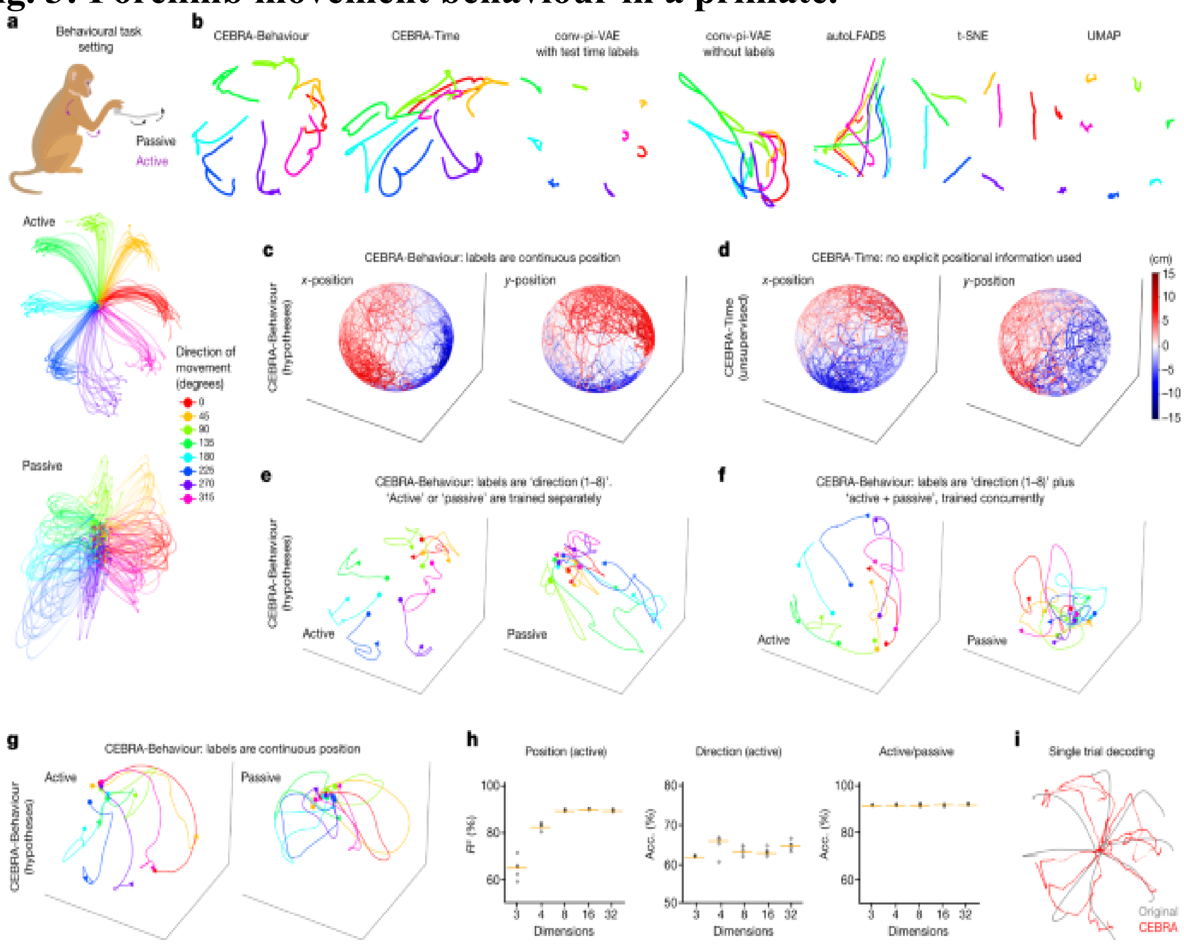
Although consistency increased, it is not a priori clear that decoding from 'pseudosubjects' would be equally good because there could be session- or animal-specific information that is lost in pseudodecoding (because decoding is usually performed within the session). Alternatively, if this joint latent space was as high performance as the single subject, that would suggest that CEBRA is able to produce robust latent spaces across subjects. Indeed, we find no loss in decoding performance (Extended Data Fig. 7c).

It is also possible to rapidly decode from a new session that is unseen during training, which is an attractive setting for brain machine interface deployment. We show that, by pretraining on a subset of the subjects, we can apply and rapidly adapt CEBRA-Behaviour on unseen data (that is, it runs at 50–100 steps $s^{-1}$, and positional decoding error already decreased by 10 cm after adapting the pretrained network for one step). Lastly, we can achieve a lower error more rapidly compared with training fully on the unseen individual (Extended Data Fig. 7d). Collectively, this shows that CEBRA can rapidly produce high-performance, consistent and robust latent spaces.

# Latent dynamics during a motor task

We next consider an eight-direction ‘centre-out’ reaching task paired with electrophysiology recordings in primate somatosensory cortex (S1)[34] (Fig. 3a). The monkey performed many active movements, and in a subset of trials experienced randomized bumps that caused passive limb movement. CEBRA produced highly informative visualizations of the data compared with other methods (Fig. 3b), and CEBRA-Behaviour can be used to test the encoding properties of S1. Using either position or time information showed embeddings with clear positional encoding (Fig. 3c,d and Extended Data Fig. 8a–c).

**Fig. 3: Forelimb movement behaviour in a primate.**




**a**, The monkey makes either active or passive movements in eight directions. Data derived from area 2 of primary S1 from Chowdhury et al.[34]. Cartoon from scidraw.io. **b**, Embeddings of active trials generated with 4D CEBRA-Behaviour, 4D CEBRA-Time, 4D autoLFADS, 4D conv-pi-VAE variants, 2D *t*-SNE and 2D UMAP. The embeddings of trials ($n = 364$) for

each direction are post hoc averaged. **c**, CEBRA-Behaviour trained with (*x*,*y*) position of the hand. Left, colour coded to *x* position; right, colour coded to *y* position. **d**, CEBRA-Time with no external behaviour variables. Colour coded as in **c**. **e**, CEBRA-Behaviour embedding trained using a 4D latent space, with discrete target direction as behaviour labels, trained and plotted separately for active and passive trials. **f**, CEBRA-Behaviour embedding trained using a 4D latent space, with discrete target direction and active and passive trials as auxiliary labels plotted separately, active versus passive trials. **g**, CEBRA-Behaviour embedding trained with a 4D latent space using active and passive trials with continuous (*x*,*y*) position as auxiliary labels plotted separately, active versus passive trials. The trajectory of each direction is averaged across trials ($n = 18$–30 each, per direction) over time. Each trajectory represents 600 ms from −100 ms before the start of the movement. **h**, Left to right, decoding performance of: position using CEBRA-Behaviour trained with (*x*,*y*) position (active trials); target direction using CEBRA-Behaviour trained with target direction (active trials); and active versus passive accuracy (Acc.) using CEBRA-Behaviour trained with both active and passive movements. In each case we trained and evaluated five seeds, represented by black dots; orange line represents the median. **i**, Decoded trajectory of hand position using CEBRA-Behaviour trained on active trial with (*x*,*y*) position of the hand. The grey line represents a true trajectory and the red line represents a decoded trajectory.

To test how directional information and active versus passive movements influence population dynamics in S1 (refs. [34,35,36]), we trained embedding spaces with directional information and then either separated the trials into active and passive for training (Fig. 3e) or trained jointly and post hoc plotted separately (Fig. 3f). We find striking similarities suggesting that active versus passive strongly influences the neural latent space: the embeddings for active trials show a clear start and stop whereas for passive trials they show a continuous trajectory through the embedding, independently of how they are trained. This finding is confirmed in embeddings that used only the continuous position of the end effector as the behavioural label (Fig. 3g). Notably, direction is a less prominent feature (Fig. 3g) although they are entangled parameters in this task.

As the position and active or passive trial type appear robust in the embeddings, we further explored the decodability of the embeddings. Both

position and trial type were readily decodable from 8D+ embeddings with a kNN decoder trained on position only, but directional information was not as decodable (Fig. 3h). Here too, the loss function value is informative for goodness of fit during hypothesis testing (Extended Data Fig. 8d–f). Notably, we could recover the hand trajectory with $R^2 = 88\%$ (concatenated across 26 held-out test trials; Fig. 3i) using a 16D CEBRA-Behaviour model trained on position (Fig. 3i). For comparison, an L1 regression using all neurons achieved $R^2 = 74\%$ and 16D conv-pi-VAE achieved $R^2 = 82\%$. We also tested CEBRA on an additional monkey dataset (mc-maze) presented in the Neural Latent Benchmark[37], in which it achieved state-of-the-art behaviour (velocity) decoding performance (Extended Data Fig. 8).

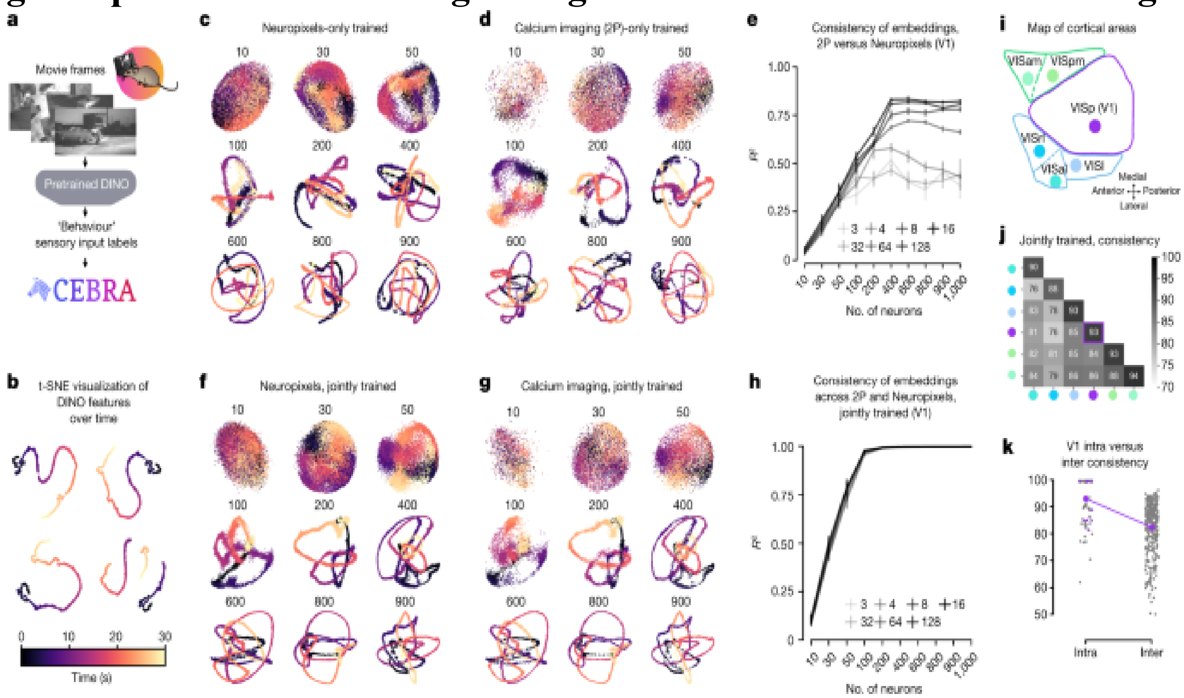
# Consistent embeddings across modalities

Although CEBRA is agnostic to the recording modality of neural data, do different modalities produce similar latent embeddings? Understanding the relationship of calcium signalling and electrophysiology is a debated topic, yet an underlying assumption is that they inherently represent related, yet not identical, information. Although there is a wealth of excellent tools aimed at inferring spike trains from calcium data, currently the pseudo-$R^2$ of algorithms on paired spiking and calcium data tops out at around 0.6 (ref. [38]). Nonetheless, it is clear that recording with either modality has led to similar global conclusions—for example, grid cells can be uncovered in spiking or calcium signals[33,39], reward prediction errors can be found in dopamine neurons across species and recording modalities[40,41,42], and visual cortex shows orientation tuning across species and modalities[43,44,45].

We aimed to formally study whether CEBRA could capture the same neural population dynamics either from spikes or calcium imaging. We utilized a dataset from the Allen Brain Observatory where mice passively watched three videos repeatedly. We focused on paired data from ten repeats of 'Natural Movie 1' where neural data were recorded with either Neuropixels (NP) probes or calcium imaging with a two-photon (2P) microscope (from separate mice)[46,47]. Note that, although the data we have considered thus far have goal-driven actions of the animals (such as running down a linear track

or reaching for targets), this visual cortex dataset was collected during passive viewing (Fig. 4a).

**Fig. 4: Spikes and calcium signalling show similar CEBRA embeddings.**




**a**, CEBRA-Behaviour can use frame-by-frame video features as a label of sensory input to extract the neural latent space of the visual cortex of mice watching a video. Cartoon from scidraw.io. **b**, *t*-SNE visualization of the DINO features of video frames from four different DINO configurations (latent size, model size), all showing continuous evolution of video frames over time. **c**,**d**, Visualization of trained 8D latent CEBRA-Behaviour embeddings with Neuropixels (NP) data (**c**) or calcium imaging (2P) (**d**). Numbers above each embedding indicate neurons subsampled from the multi-session concatenated dataset. Colour map as in **b**. **e**, Linear consistency between embeddings trained with either calcium imaging or Neuropixels data ($n = 10$–1,000 neurons, across $n = 5$ shuffles of neural data; mean values ± s.e.m.). **f**,**g**, Visualization of CEBRA-Behaviour embedding (8D) trained jointly with Neuropixels (**f**) and calcium imaging (**g**). Colour map as in **b**. **h**, Linear consistency between embeddings of calcium imaging and Neuropixels trained jointly using a multi-session CEBRA model ($n = 10$–1000 neurons, across $n = 5$ shuffles of neural data; mean values ± s.e.m.). **i**, Diagram of mouse primary visual cortex (V1, VISp) and higher visual areas. **j**, CEBRA-Behaviour 32D model jointly

trained with 2P + NP incorporating 400 neurons, followed by measurement of consistency within or across areas (2P versus NP) across two unique sets of disjoint neurons for three seeds and averaged. **k**, Models trained as in **h**, with intra-V1 consistency measurement versus all interarea versus V1 comparison. Purple dots indicate mean of V1 intra-V1 consistency (across $n = 120$ runs) and inter-V1 consistency ($n = 120$ runs). Intra-V1 consistency is significantly higher than interarea consistency (one-sided Welch's *t*-test, $t(12.30) = 4.55$, $P = 0.00019$).

We used the video features as 'behaviour' labels by extracting high-level visual features from the video on a frame-by-frame basis with DINO, a powerful vision transformer model[48]. These were then used to sample the neural data with feature-labels (Fig. 4b). Next, we used either Neuropixels or 2P data (each with multi-session training) to generate (from 8D to 128D) latent spaces from varying numbers of neurons recorded from primary visual cortex (V1) (Fig. 4c,d). Visualization of CEBRA-Behaviour showed trajectories that smoothly capture the video of either modality with an increasing number of neurons. This is reflected quantitatively in the consistency metric (Fig. 4e). Strikingly, CEBRA-Time efficiently captured the ten repeats of the video (Extended Data Fig. 9), which was not captured by other methods. This result demonstrates that there is a highly consistent latent space independent of the recording method.

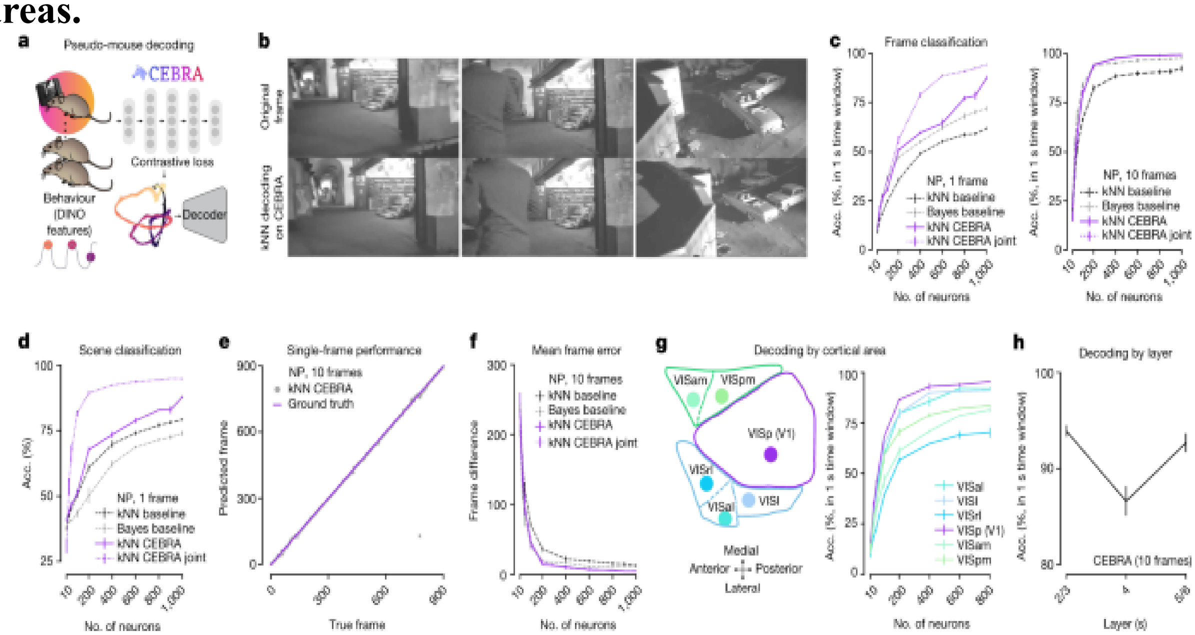
Next, we stacked neurons from different mice and modalities and then sampled random subsets of V1 neurons to construct a pseudomouse. We did not find that joint training lowered consistency within modality (Extended Data Fig. 10a,b) and, overall, we found considerable improvement in consistency with joint training (Fig. 4f–h).

Using CEBRA-Behaviour or -Time, we trained models on five higher visual areas and measured consistency with and without joint training, and within or across areas. Our results show that, with joint training, intra-area consistency is higher compared with other areas (Fig. 4i–k), suggesting that CEBRA is not removing biological differences across areas, which have known differences in decodability and feature representations[49,50]. Moreover, we tested within modality and find a similar effect for CEBRA-Behaviour and -Time within recording modality (Extended Data Fig. 10c–f).

# Decoding of natural videos from cortex

We performed V1 decoding analysis using CEBRA models that are either joint-modality trained, single-modality trained or with a baseline population vector paired with a simple kNN or naive Bayes decoder. We aimed to determine whether we could decode, on a frame-by-frame basis, the natural video watched by the mice. We used the final video repeat as a held-out test set and nine repeats as the training set. We achieved greater than 95% decoding accuracy, which is significantly better than baseline decoding methods (naive Bayes or kNN) for Neuropixels recordings, and joint-training CEBRA outperformed Neuropixels-only CEBRA-based training (single frame: one-way ANOVA, $F(3,197) = 5.88$, $P = 0.0007$; Supplementary Tables 3–5, Fig. 5a–d and Extended Data Fig. 10g,h). Accuracy was defined by either the fraction of correct frames within a 1 s window or identification of the correct scene. Frame-by-frame results also showed reduced frame ID errors (one-way ANOVA, $F(3,16) = 20.22$, $P = 1.09 \times 10^{-5}$, $n = 1{,}000$ neurons; Supplementary Table 6), which can be seen in Fig. 5e,f, Extended Data Fig. 10i and Supplementary Video 2. The DINO features themselves did not drive performance, because shuffling of features showed poor decoding (Extended Data Fig. 10j).

**Fig. 5: Decoding of natural video features from mouse visual cortical areas.**



**a**, Schematic of the CEBRA encoder and kNN (or naive Bayes) decoder. **b**, Examples of original frames (top row) and frames decoded from CEBRA embedding of V1 calcium recording using kNN decoding (bottom row). The last repeat among ten was used as the held-out test. **c**, Decoding accuracy measured by considering a predicted frame being within 1 s of the true frame as a correct prediction using CEBRA (NP only), jointly trained (2P + NP) or a baseline population vector plus kNN or naive Bayes decoder using either a one-frame (33 ms) receptive field (left) or ten frames (330 ms) (right); results shown for Neuropixels dataset (V1 data); for each neuron number we have $n = 5$ shuffles, mean ± s.e.m. **d**, Decoding accuracy measured by correct scene prediction using either CEBRA (NP only), jointly trained (2P + NP) or baseline population vector plus kNN or Bayes decoder using a one-frame (33 ms) receptive field (V1 data); $n = 5$ shuffles per neuron number, mean ± s.e.m. **e**, Single-frame ground truth frame ID versus predicted frame ID for Neuropixels using a CEBRA-Behaviour model trained with a 330 ms receptive field (1,000 V1 neurons across mice used). **f**, Mean absolute error of the correct frame index; shown for baseline and CEBRA models as computed in **c**–**e**. **g**, Diagram of the cortical areas considered and decoding performance from CEBRA (NP only), ten-frame receptive field; $n = 3$ shuffles for each area and number of neurons, mean ± s.e.m. **h**, V1 decoding performance versus layer category using 900 neurons with a 330 ms receptive field CEBRA-Behaviour model; $n = 5$ shuffles for each layer, mean ± s.e.m.

Lastly, we tested decoding from other higher visual areas using DINO features. Overall, decoding from V1 had the highest performance and VISrl the lowest (Fig. 5g and Extended Data Fig. 10k). Given the high decoding performance of CEBRA, we tested whether there was a particular V1 layer that was most informative. We leveraged CEBRA-Behaviour by training models on each category and found that layers 2/3 and 5/6 showed significantly higher decoding performance compared with layer 4 (one-way ANOVA, $F(2,12) = 9.88$, $P = 0.003$; Fig. 5h). Given the known cortical connectivity, this suggests that the nonthalamic input layers render frame information more explicit, perhaps via feedback or predictive processing.

## Discussion

CEBRA is a nonlinear dimensionality reduction method newly developed to explicitly leverage auxiliary (behaviour) labels and/or time to discover latent features in time series data—in this case, latent neural embeddings. The unique property of CEBRA is the extension and generalization of the standard InfoNCE objective by introduction of a variety of different sampling strategies tuned for usage of the algorithm in the experimental sciences and for analysis of time series datasets, and it can also be used for supervised and self-supervised analysis, thereby directly facilitating hypothesis- and discovery-driven science. It produces both consistent embeddings across subjects (thus showing common structure) and can find the dimensionality of neural spaces that are topologically robust. Although there remains a gap in our understanding of how these latent spaces map to neural-level computations, we believe this tool provides an advance in our ability to map behaviour to neural populations. Moreover, because pretrained CEBRA models can be used for decoding in new animals within tens of steps (milliseconds), we can thereby obtain equal or better performance compared with training on the unseen animal alone.

Dimensionality reduction is often tightly linked to data visualization, and here we make an empirical argument that ultimately this is useful only when obtaining consistent results and discovering robust features. Unsupervised *t*-SNE and UMAP are examples of algorithms widely used in life sciences for discovery-based analysis. However, they do not leverage time and, for neural recordings, this is always available and can be used. Even more critical is that concatenation of data from different animals can lead to shifted clusters with *t*-SNE or UMAP due to inherent small changes across animals or in how the data were collected. CEBRA allows the user to remove this unwanted variance and discover robust latents that are invariant to animal ID, sessions or any-other-user-defined nuisance variable. Collectively we believe that CEBRA will become a complement to (or replacement for) these methods such that, at minimum, the structure of time in the neural code is leveraged and robustness is prioritized.

# Methods

## Datasets

## Artificial spiking dataset

The synthetic spiking data used for benchmarking in Fig. 1 were adopted from Zhou and Wei[5]. The continuous 1D behaviour variable $c\in [0,2\pi )$ was sampled uniformly in the interval $[0,2\pi )$. The true 2D latent variable ${\bf{z}}\in {{\mathbb{R}}}^{2}$ was then sampled from a Gaussian distribution ${\mathscr{N}}\left(\mu \left(c\right),\varSigma \left(c\right)\right)$ with mean $\mu \left(c\right)={\left(c,2\sin c\right)}^{\top }$ and covariance $\varSigma \left(c\right)={\rm{diag}}\left(0.6-0.3\left|\sin c\right|,0.3\left|\sin c\right|\right)$. After sampling, the 2D latent variable ${\bf{z}}$ was mapped to the spiking rates of 100 neurons by the application of four randomly initialized RealNVP[51] blocks. Poisson noise was then applied to map firing rates onto spike counts. The final dataset consisted of $1.5 \times 10^4$ data points for 100 neurons ([number of samples, number of neurons]) and was split into train (80%) and validation (20%) sets. We quantified consistency across the entire dataset for all methods. Additional synthetic data, presented in Extended Data Fig. 1, were generated by varying noise distribution in the above generative process. Beside Poisson noise, we used additive truncated ([0,1000]) Gaussian noise with s.d. = 1 and additive uniform noise defined in [0,2], which was applied to the spiking rate. We also adapted Poisson spiking by simulating neurons with a refractory period. For this, we scaled the spiking rates to an average of 110 Hz. We sampled interspike intervals from an exponential distribution with the given rate and added a refractory period of 10 ms.

## Rat hippocampus dataset

We used the dataset presented in Grosmark and Buzsáki[26]. In brief, bilaterally implanted silicon probes recorded multicellular electrophysiological data from CA1 hippocampus areas from each of four male Long–Evans rats. During a given session, each rat independently ran on a 1.6-m-long linear track where they were rewarded with water at each end of the track. The numbers of recorded putative pyramidal neurons for each rat ranged between 48 and 120. Here, we processed the data as in Zhou and Wei[5]. Specifically, the spikes were binned into 25 ms time windows. The position and running direction (left or right) of the rat were encoded into

a 3D vector, which consisted of the continuous position value and two binary values indicating right or left direction. Recordings from each rat were parsed into trials (a round trip from one end of the track as a trial) and then split into train, validation and test sets with a $k = 3$ nested cross-validation scheme for the decoding task.

### Macaque dataset

We used the dataset presented in Chowdhury et al.[34] In brief, electrophysiological recordings were performed in Area 2 of somatosensory cortex (S1) in a rhesus macaque (monkey) during a centre-out reaching task with a manipulandum. Specifically, the monkey performed an eight-direction reaching task in which on 50% of trials it actively made centre-out movements to a presented target. The remaining trials were 'passive' trials in which an unexpected 2 Newton force bump was given to the manipulandum towards one of the eight target directions during a holding period. The trials were aligned as in Pei et al.[37], and we used the data for −100 and 500 ms from movement onset. We used 1 ms time bins and convolved the data using a Gaussian kernel with s.d. = 40 ms.

### Mouse visual cortex datasets

We utilized the Allen Institute two-photon calcium imaging and Neuropixels data recorded from five mouse visual and higher visual cortical areas (VISp, VISl, VISal, VISam, VISpm and VISrl) during presentation of a monochrome video with 30 Hz frame rate, as presented previously[46,47,52]. For calcium imaging (2P) we used the processed dataset from Vries et al.[46] with a sampling rate of 30 Hz, aligned to the video frames. We considered the recordings from excitatory neurons (Emx1-IRES-Cre, Slc17a7-IRES2-Cre, Cux2-CreERT2, Rorb-IRES2-Cre, Scnn1a-Tg3-Cre, Nr5a1-Cre, Rbp4-Cre_KL100, Fezf2-CreER and Tlx3-Cre_PL56) in the 'Visual Coding-2P' dataset. Ten repeats of the first video (Movie 1) were shown in all session types (A, B and C) for each mouse and we used those neurons that were recorded in all three session types, found via cell registration[46]. The Neuropixels recordings were obtained from the 'Brain Observatory 1.1' dataset[47]. We used the preprocessed spike timings and binned them to a

sampling frequency of 120 Hz, aligned with the video timestamps (exactly four bins aligned with each frame). The dataset contains recordings for ten repeats, and we used the same video (Movie 1) that was used for the 2P recordings. For analysis of consistency across the visual cortical areas we used a disjoint set of neurons for each seed, to avoid higher intraconsistency due to overlapping neuron identities. We made three disjoint sets of neurons by considering only neurons from session A (for 2P data) and nonoverlapping random sampling for each seed.

## CEBRA model framework

### Notation

We will use $\mathbf{x}$,$\mathbf{y}$ as general placeholder variables and denote the multidimensional, time-varying signal as $\mathbf{s}_t$, parameterized by time $t$. The multidimensional, continuous context variable $\mathbf{c}_t$ contains additional information about the experimental condition and additional recordings, similar to the discrete categorical variable $k_t$.

The exact composition of $\mathbf{s}$, $\mathbf{c}$ and $k$ depends on the experimental context. CEBRA is agnostic to exact signal types; with the default parameterizations, $\mathbf{s}_t$ and $\mathbf{c}_t$ can have up to an order of hundreds or thousands of dimensions. For even higher-dimensional datasets (for example, raw video, audio and so on) other optimized deep learning tools can be used for feature extraction before the application of CEBRA.

### Applicable problem setup

We refer to ${\bf{x}}\in X$ as the reference sample and to ${\bf{y}}\in Y$ as a corresponding positive or negative sample. Together, ($\mathbf{x}$, $\mathbf{y}$) form a positive or negative pair based on the distribution from which $\mathbf{y}$ is sampled. We denote the distribution and density function of $\mathbf{x}$ as $p(\mathbf{x})$, the conditional distribution and density of the positive sample $\mathbf{y}$ given $\mathbf{x}$ as $p({\bf{y}}{\rm{| }}{\bf{x}})$ and the conditional distribution and density of the negative sample $\mathbf{y}$ given $\mathbf{x}$ as $q\left({\bf{y}}{\rm{| }}{\bf{x}}\right)$.

After sampling—and irrespective of whether we are considering a positive or negative pair—samples ${\bf{x}}\in {{\mathbb{R}}}^{D}$ and ${\bf{y}}\in {{\mathbb{R}}}^{{D}^{{\prime} }}$ are encoded by feature extractors ${\bf{f}}:X\mapsto Z$ and ${{\bf{f}}}^{{\prime} }:Y\mapsto Z$. The feature extractors map both samples from signal space $X\subseteq {{\mathbb{R}}}^{D},Y\subseteq {{\mathbb{R}}}^{{D}^{{\prime} }}$ into a common embedding space $Z\subseteq {{\mathbb{R}}}^{E}$. The design and parameterization of the feature extractor are chosen by the user of the algorithm. Note that spaces $X$ and $Y$ and their corresponding feature extractors can be the same (which is the case for single-session experiments in this work), but that this is not a strict requirement within the CEBRA framework (for example, in multi-session training across animals or modalities, $X$ and $Y$ are selected as signals from different mice or modalities, respectively). It is also possible to include the context variable (for example, behaviour) into $X$, or to set $\mathbf{x}$ to the context variable and $\mathbf{y}$ to the signal variable.

Given two encoded samples, a similarity measure $\varphi :Z\times Z\mapsto {\mathbb{R}}$ assigns a score to a pair of embeddings. The similarity measure needs to assign a higher score to more similar pairs of points, and to have an upper bound. For this work we consider the dot product between normalized feature vectors, $\varphi ({\bf{z}},{{\bf{z}}}^{{\prime} })={{\bf{z}}}^{{\rm{\top }}}{{\bf{z}}}^{{\prime} }/\tau$, in most analyses (latents on a hypersphere) or the negative mean squared error, $\varphi ({\bf{z}},{{\bf{z}}}^{{\prime} })=-\,\parallel {\bf{z}}-{{\bf{z}}}^{{\prime} }{\parallel }^{2}/\tau$ (latents in Euclidean space). Both metrics can be scaled by a temperature parameter $\tau$ that is either fixed or jointly learned with the network. Other $L_p$ norms and other similarity metrics, or even a trainable neural network (a so-called projection head commonly used in contrastive learning algorithms[14,22]), are possible choices within the CEBRA software package. The exact choice of $\phi$ shapes the properties of the embedding space and encodes assumptions about distributions $p$ and $q$.

The technique requires paired data recordings—for example, as is common in aligned time series. The signal $\mathbf{s}_t$, continuous context $\mathbf{c}_t$ and discrete context $\mathrm{k}_t$ are synced in their time point $t$. How the reference, positive and

negative samples are constructed from these available signals is a configuration choice made by the algorithm user, and depends on the scientific question under investigation.

**Optimization**

Given the feature encoders **f** and **f′** for the different sample types, as well as the similarity measure $\phi$, we introduce the shorthand $\psi \left({\bf{x}},{\bf{y}}\right)=\varphi \left({\bf{f}}\left({\bf{x}}\right),{{\bf{f}}}^{{\prime} }\left({\bf{y}}\right)\right)$. The objective function can then be compactly written as:

$$\mathop{\int }\limits_{{\bf{x}}\in {\rm{X}}}{\rm{d}}{\bf{x}}{\rm{p}}({\bf{x}})\,[\,-\mathop{\int }\limits_{{\bf{y}}\in {\rm{Y}}}{\rm{d}}{\bf{y}}{\rm{p}}({\bf{y}}|{\bf{x}})\psi ({\bf{x}},{\bf{y}})+\log \mathop{\int }\limits_{{\bf{y}}\in {\rm{Y}}}{\rm{d}}{\bf{y}}{\rm{q}}({\bf{y}}|{\bf{x}}){{\rm{e}}}^{\psi ({\bf{x}},{\bf{y}})}].$$
(1)

We approximate this objective by drawing a single positive example $\mathbf{y}_+$, and multiple negative examples $\mathbf{y}_i$ from the distributions outlined above, and minimize the loss function

$$\mathop{{\mathbb{E}}}\limits_{\begin{array}{c}{\bf{x}}\sim p({\bf{x}}),{{\bf{y}}}_{+}\sim p({\bf{y}}|{\bf{x}})\\ {{\bf{y}}}_{1},\ldots ,{{\bf{y}}}_{n}\sim q({\bf{y}}|{\bf{x}})\end{array}}\,[-\psi ({\bf{x}},{{\bf{y}}}_{+})+\,\log \mathop{\sum }\limits_{i=1}^{n}{e}^{\psi ({\bf{x}},{{\bf{y}}}_{i})}],$$
(2)

with a gradient-based optimization algorithm. The number of negative samples is a hyperparameter of the algorithm, and larger batch sizes are generally preferable.

For sufficiently small datasets, as used in this paper, both positive and negative samples are drawn from all available samples in the dataset. This is

in contrast to the common practice in many contrastive learning frameworks in which a minibatch of samples is drawn first, which are then grouped into positive and negative pairs. Allowing access to the whole dataset to form pairs gives a better approximation of the respective distributions $p\left({\bf{y}}{\rm{| }}{\bf{x}}\right)$ and $q\left({\bf{y}}{\rm{| }}{\bf{x}}\right)$, and considerably improves the quality of the obtained embeddings. If the dataset is sufficiently small to fit into the memory, CEBRA can be optimized with batch gradient descent—that is, using the whole dataset at each optimizer step.

**Goodness of fit**

Comparing the loss value—at both absolute and relative values across models at the same point in training time— can be used to determine goodness of fit. In practical terms, this means that one can find which hypothesis best fits one's data, in the case of using CEBRA-Behaviour. Specifically, let us denote the objective in equation(1) as $L_{\mathrm{asympt}}$ and its approximation in equation (2) with a batch size of $n$ as $L_n$. In the limit of many samples, the objective converges up to a constant, ${L}_{{\rm{asympt}}}={{\rm{lim}}}_{n\to \infty }\left[{L}_{n}-\log n\right]$ (Supplementaty Note 2 and ref. [53]).

The objective also has two trivial solutions: the first is obtained for a constant $\psi \left({\bf{x}},{\bf{y}}\right)=\psi$, which yields a value of $L_n = \log n$. This solution can be obtained when the labels are not related to the signal (e.g., with shuffled labels). It is typically not obtained during regular training because the network is initialized randomly, causing the initial embedding points to be randomly distributed in space.

If the embedding points are distributed uniformly in space and $\phi$ is selected such that ${\mathbb{E}}\left[\varphi \left({\bf{x}},{\bf{y}}\right)\right]=0$, we will also get a value that is approximately $L_n = \log n$. This value can be readily estimated by computing $\varphi \left({\bf{u}},{\bf{v}}\right)$ for randomly distributed points.

The minimizer of equation (1) is also clearly defined as ${-D}_{{\rm{KL}}}\left(p\parallel q\right)$ and depends on the positive and

negative distribution. For discovery-driven (time contrastive) learning, this value is impossible to estimate because it would require access to the underlying conditional distribution of the latents. However, for training with predefined positive and negative distributions, this quantity can be again numerically estimated.

Interesting values of the loss function when fitting a CEBRA model are therefore

$$-{D}_{{\rm{KL}}}\left(p\parallel q\right)\le {L}_{n}-\log n\,\le 0$$
(3)

where $L_n - \log n$ is the goodness of fit (lower is better) of the CEBRA model. Note that the metric is independent of the batch size used for training.

### Sampling

Selection of the sampling scheme is CEBRA's key feature in regard to adapting embedding spaces to different datasets and recording setups. The conditional distributions $p\left({\bf{y}}{\rm{| }}{\bf{x}}\right)$ for positive samples and $q\left({\bf{y}}{\rm{| }}{\bf{x}}\right)$ for negative samples, as well as the marginal distribution $p(\mathbf{x})$ for reference samples, are specified by the user. CEBRA offers a set of predefined sampling techniques but customized variants can be specified to implement additional, domain-specific distributions. This form of training allows the use of context variables to shape the properties of the embedding space, as outlined in the graphical model in Supplementary Note 1.

Through the choice of sampling technique, various use cases can be built into the algorithm. For instance, by forcing positive and negative distributions to sample uniformly across a factor, the model will become invariant to this factor because its inclusion would yield a suboptimal value of the objective function.

When considering different sampling mechanisms we distinguish between single- and multi-session datasets: a single-session dataset consists of samples $\mathbf{s}_t$ associated to one or more context variables $\mathbf{c}_t$ and/or $k_t$. These

context variables allow imposition of the structure on the marginal and conditional distribution used for obtaining the embedding. Multi-session datasets consist of multiple, single-session datasets. The dimension of context variables $\mathbf{c}_t$ and/or $k_t$ must be shared across all sessions whereas the dimension of the signal $\mathbf{s}_t$ can vary. In such a setting, CEBRA allows learning of a shared embedding space for signals from all sessions.

For single-session datasets, sampling is done in two steps. First, based on a specified 'index' (the user-defined context variable $\mathbf{c}_t$ and/or $k_t$), locations $t$ are sampled for reference, positive and negative samples. The algorithm differentiates between categorical ($k$) and continuous ($\mathbf{c}$) variables for this purpose.

In the simplest case, negative sampling ($q$) returns a random sample from the empirical distribution by returning a randomly chosen index from the dataset. Optionally, with a categorical context variable ${k}_{t}\in \left[K\right]$, negative sampling can be performed to approximate a uniform distribution of samples over this context variable. If this is performed for both negative and positive samples, the resulting embedding will become invariant with respect to the variable $k_t$. Sampling is performed in this case by computing the cumulative histogram of $k_t$ and sampling uniformly over $k$ using the transformation theorem for probability densities.

For positive pairs, different options exist based on the availability of continuous and discrete context variables. For a discrete context variable ${k}_{t}\in \left[K\right]$ with $K$ possible values, sampling from the conditional distribution is done by filtering the whole dataset for the value $k_t$ of the reference sample, and uniformly selecting a positive sample with the same value. For a continuous context variable $\mathbf{c}_t$ we can use a set of time offsets $\varDelta$ to specify the distribution. Given the time offsets, the empirical distribution $P\left({{\bf{c}}}_{t+\tau }\,{\rm{| }}\,{{\bf{c}}}_{t}\right)$ for a particular choice of $\tau \in \varDelta$ can be computed from the dataset: we build up a set $D=\{t\in \left[T\right],\tau \in \varDelta :{{\bf{c}}}_{t+\tau }-{{\bf{c}}}_{t}\}$, sample a $\mathbf{d}$ uniformly from $D$ and obtain the sample that is closest to the reference sample's context variable modified by this distance ($\mathbf{c} + \mathbf{d}$) from the dataset. It is possible to combine a

continuous variable $\mathbf{c}_t$ with a categorical variable $k_t$ for mixed sampling. On top of the continual sampling step above, it is ensured that both samples in the positive pair share the same value of $k_t$.

It is crucial that the context samples $\mathbf{c}$ and the norm used in the algorithm match in some way; for simple context variables with predictable conditional distributions (for example, a 1D or 2D position of a moving animal, which can most probably be well described by a Gaussian conditional distribution based on the previous sample), the positive sample distribution can also be specified directly, for example, as a normal distribution centred around $\mathbf{c}_t$. An additional alternative is to use CEBRA also to preprocess the original context samples $\mathbf{c}$ and use the embedded context samples with the metric used for CEBRA training. This scheme is especially useful for higher-dimensional behavioural data, or even for complex inputs such as video.

We next consider the multi-session case in which signals ${{\bf{s}}}_{t}^{\left(i\right)}\in {{\mathbb{R}}}^{{n}_{i}}$ come from $N$ different sessions $i\in \left[N\right]$ with session-dependent dimensionality $n_i$. Importantly, the corresponding continuous context variables ${{\bf{c}}}_{t}^{\left(i\right)}\in {{\mathbb{R}}}^{m}$ share the same dimensionality $m$, which makes it possible to relate samples across sessions. The multi-session setup is similar to mixed-session sampling (if we treat the session ID as a categorical variable ${k}_{t}^{(i)}:\,=i$ for all time steps $t$ in session $i$). The conditional distribution for both negative and positive pairs is uniformly sampled across sessions, irrespective of session length. Multi-session mixed or discrete sampling can be implemented analogously.

CEBRA is sufficiently flexible to incorporate more specialized sampling schemes beyond those outlined above. For instance, mixed single-session sampling could be extended additionally to incorporate a dimension to which the algorithm should become invariant; this would add an additional step of uniform sampling with regard to this desired discrete variable (for example, via ancestral sampling).

**Choice of reference, positive and negative samples**

Depending on the exact application, the contrastive learning step can be performed by explicitly including or excluding the context variable. The reference sample $\mathbf{x}$ can contain information from the signal $\mathbf{s}_t$, but also from the experimental conditions, behavioural recordings or other context variables. The positive and negative samples $\mathbf{y}$ are set to the signal variable $\mathbf{s}_t$.

### Theoretical guarantees for linear identifiability of CEBRA models

Identifiability describes the property of an algorithm to give a consistent estimate for the model parameters given that the data distributions match. We here apply the relaxed notion of linear identifiability that was previously discussed and used[13,14]. After training two encoder models $\mathbf{f}$ and $\mathbf{f}'$, the models are linearly identifiable if $\mathbf{f}(\mathbf{x}) = \mathbf{L}\mathrm{f}(\mathbf{x})$, where $\mathbf{L}$ is a linear map.

When applying CEBRA, three cases are of potential interest. (1) When applying discovery-driven CEBRA, will two models estimated on comparable experimental data agree in their inferred representation? (2) Under which assumptions about the data will we be able to discover the true latent distribution? (3) In the hypothesis-driven or hybrid application of CEBRA, is the algorithm guaranteed to give a meaningful (nonstandard) latent space when we can find signal within the data?

For the first case, we note that the CEBRA objective with a cosine similarity metric follows the canonical discriminative form for which Roeder et al.[13] showed linear identifiability: for sufficiently diverse datasets, two CEBRA models trained to convergence on the same dataset will be consistent up to linear transformations. Note that the consistency of CEBRA is independent of the exact data distribution: it is merely required that the embeddings of reference samples across multiple positive pairs, and the embeddings of negative samples across multiple negative pairs, vary in sufficiently numerous linearly independent directions. Alternatively, we can derive linear identifiability from assumptions about data distribution: if the ground truth latents are sufficiently diverse (that is, vary in all latent directions under distributions $p$ and $q$), and the model is sufficiently parameterized to fit the data, we will also obtain consistency up to a linear transformation. See Supplementary Note 2 for a full formal discussion and proof.

For the second case, additional assumptions are required regarding the exact form of data-generating distributions. Within the scope of this work we consider ground truth latents distributed on the hypersphere or Euclidean space. The metric then needs to match assumptions about the variation of ground truth latents over time. In discovery-driven CEBRA, using the dot product as the similarity measure then encodes the assumption that latents vary according to a von Mises–Fisher distribution whereas the (negative) mean squared error encodes an assumption that latents vary according to a normal distribution. More broadly, if we assume that the latents have a uniform marginal distribution (which can be ensured by designing unbiased experiments), the similarity measure should be chosen as the log-likelihood of conditional distribution over time. In this case, CEBRA identifies the latents up to an affine transformation (in the most general case).

This result also explains the empirically high performance of CEBRA for decoding applications: if trained for decoding (using the variable to decode for informing the conditional distribution), it is trivial to select matching conditional distributions because both quantities are directly selected by the user. CEBRA then 'identifies' the context variable up to an affine transformation.

For the third case, we are interested in hypothesis-testing capabilities. We can show that if a mapping exists between the context variable and the signal space, CEBRA will recover this relationship and yield a meaningful embedding, which is also decodable. However, if such a mapping does not exist we can show that CEBRA will not learn a structured embedding.

## CEBRA models

We chose $X = Y$ as the neural signal, with varying levels of recorded neurons and channels based on the dataset. We used three types of encoder model based on the required receptive field: a receptive field of one sample was used for the synthetic dataset experiments (Fig. 1b) and a receptive field of ten samples in all other experiments (rat, monkey, mouse) except for the Neuropixels dataset, in which a receptive field of 40 samples was used due to the fourfold higher sampling rate of the dataset.

All feature encoders were parameterized by the number of neurons (input dimension), a hidden dimension used to control model size and capacity, as well as by their output (embedding) dimension. For the model with the receptive field of one, a four-layer MLP was used. The first and second layers map their respective inputs to the hidden dimension whereas the third introduces a bottleneck and maps to half the hidden dimension. The final layer maps to the requested output dimension. For the model with a receptive field of ten, a convolutional network with five time-convolutional layers was used. The first layer had a kernel size of two, and the next three had a kernel size of three and used skip connections. The final layer had a kernel size of three and mapped hidden dimensions to the output dimension. For the model with receptive field 40, we first preprocessed the signal by concatenating a 2× downsampled version of the signal with a learnable downsample operation implemented as a convolutional layer with kernel size four and stride two, directly followed (without activation function between) by another convolutional layer with kernel size three and stride two. After these first layers, the signal was subsampled by a factor of four. Afterwards, similar to the receptive field ten model, we applied three layers with kernel size three and skip connections and a final layer with kernel size three. In all models, Gaussian error linear unit activation functions[54] were applied after each layer except the last. The feature vector was normalized after the last layer unless a mean squared error-based similarity metric was used (as shown in Extended Data Fig. 8).

Our implementation of the InfoNCE criterion received a minibatch (or the full dataset) of size $n \times d$ for each of the reference, positive and negative samples. $n$ dot-product similarities were computed between reference and positive samples and $n \times n$ dot-product similarities were computed between reference and negative samples. Similarities were scaled with the inverse of the temperature parameter $\tau$:

```
from torch import einsum, logsumexp, no_grad def info_nce(ref, pos, neg, tau = 1.0): pos_dist = einsum(“nd,nd–>n”, ref, pos)/tau neg_dist = einsum(“nd,md–>nm”, ref, neg)/tau with no_grad(): c, _ = neg_dist.max(dim=1) pos_dist = pos_dist – c.detach() neg_dist = neg_dist – c.detach() pos_loss = –pos_dist.mean() neg_loss = logsumexp(neg_dist, dim = 1).mean() return pos_loss + neg_loss
```

Alternatively, a learnable temperature can be used. For a numerically stable implementation we store the log inverse temperature $\alpha =-\,\log \tau$ as a parameter of loss function. At each step we scale the distances in loss function with $\min \left(\exp \alpha ,\,1/{\tau }_{\min }\right)$. The additional parameter $\tau_{\min}$ is a lower bound on the temperature. The inverse temperature used for scaling the distances in the loss will hence lie in $(0,1/{\tau }_{\min }]$.

**CEBRA model parameters used**

In the main figures we have used the default parameters (https://cebra.ai/docs/api.html) for fitting CEBRA unless otherwise stated in the text (such as dimension, which varied and is noted in figure legends), or below:

Synthetic data: model_architecture= ‘offset1-model-mse’, conditional= ‘delta’, delta=0.1, distance= ‘euclidean’, batch_size=512, learning_rate=1e-4.

Rat hippocampus neural data: model_architecture= ‘offset10-model’, time_offsets=10, batch_size=512.

Rat behavioural data: model_architecture= ‘offset10-model-mse’, distance= ‘euclidean’, time_offsets=10, batch_size=512.

Primate S1 neural data: model_architecture= ‘offset10-model’, time_offsets=10, batch_size=512.

Allen datasets (2P): model_architecture= ‘offset10-model’, time_offsets=10, batch_size=512.

Allen datasets (NP): model_architecture= ‘offset40-model-4x-subsample’, time_offsets=10, batch_size=512.

**CEBRA API and example usage**

The Python implementation of CEBRA is written in PyTorch[55] and NumPy[56] and provides an application programming interface (API) that is fully compatible with scikit-learn[57], a package commonly used for machine learning. This allows the use of scikit-learn tools for hyperparameter selection and downstream processing of the embeddings—for example, decoding. CEBRA can be used as a dropin replacement in existing data pipelines for algorithms such as *t*-SNE, UMAP, PCA or FastICA. Both CPU and GPU implementations are available.

Using the previously introduced notations, suppose we have a dataset containing signals $\mathbf{s}_t$, continuous context variables $\mathbf{c}_t$ and discrete context variables $k_t$ for all time steps $t$,

import numpy as np N = 500 s = np.zeros((N, 55), dtype = float) k = np.zeros((N,), dtype = int) c = np.zeros((N, 10), dtype = float)

along with a second session of data,

s2 = np.zeros((N, 75), dtype = float) c2 = np.zeros((N, 10), dtype = float)

assert c2.shape[1] == c.shape[1]:note that both the number of samples and the dimension in $\mathbf{s}'$ does not need to match $\mathbf{s}$. Session alignment leverages the fact that the second dimensions of $\mathbf{c}$ and $\mathbf{c}'$ match. With this dataset in place, different variants of CEBRA can be applied as follows:

import cebra model = cebra.CEBRA (output_dimension=8, num_hidden_units=32, batch_size=1024, learning_rate=3e-4, max_iterations=1000)

The training mode to use is determined automatically based on what combination of data is passed to the algorithm:

*# time contrastive learning model.fit(s) # discrete behaviour contrastive learning model.fit(s, k) # continuous behaviour contrastive learning model.fit(s, c) # mixed behaviour contrastive learning model.fit(s, c, k) # multi-session training model.fit([s, s2], [c, c2]) # adapt to new session model.fit(s, c) model.fit(s2, c2, adapt = True)*

Because CEBRA is a parametric method training a neural network internally, it is possible to embed new data points after fitting the model:

s_test = np.zeros((N, 55), dtype=float) # obtain and plot embedding z = model.transform(s_test) plt.scatter(z[:, 0], z[:, 1]) plt.show()

Besides this simple-to-use API for end users, our implementation of CEBRA is a modular software library that includes a plugin system, allowing more advanced users to readily add additional model implementations, similarity functions, datasets and data loaders and distributions for sampling positive and negative pairs.

## Consistency of embeddings across runs, subjects, sessions, recording modalities and areas

To measure the consistency of the embeddings we used the $R^2$ score of linear regression (including an intercept) between embeddings from different subjects (or sessions). Secondly, pi-VAE, which we benchmarked and improved (Extended Data Fig. 1), demonstrated a theoretical guarantee that it can reconstruct the true latent space up to an affine transformation. Across runs, we measured the $R^2$ score of linear regression between embeddings across ten runs of the algorithms, yielding 90 comparisons. These runs were done with the same hyperparameters, model and training setup.

For the rat hippocampus data, the numbers of neurons recorded were different across subjects. The behaviour setting was the same: the rats moved along a 1.6-meter-long track and, for analysis, behaviour data were binned into 100 bins of equal size for each direction (leftwards, rightwards). We computed averaged feature vectors for each bin by averaging all normalized CEBRA embeddings for a given bin and renormalized the average to lie on the hypersphere. If a bin did not contain any sample, it was filled by samples from the two adjacent bins. CEBRA was trained with latent dimension three (the minimum) such that it was constrained to lie only on a two-sphere (making this ‘3D’ space equivalent to a 2D Euclidean space). All other methods were trained with two latent dimensions in Euclidean space. Note that $n + 1$ dimensions of CEBRA are equivalent to $n$ dimensions of other methods that we compared, because the feature space of

CEBRA is normalized (that is, the feature vectors are normalized to have unit length).

For Allen visual data in which the number of behavioural data points is the same across different sessions (that is, fixed length of video stimuli), we directly computed the $R^2$ score of linear regression between embeddings from different sessions and modalities. We surveyed three, four, eight, 32, 64 and 128 latent dimensions with CEBRA.

To compare the consistency of embeddings between or within the areas considered, we computed intra- and interarea consistency within the same recording modality (2P or NP). Within the same modality we sampled 400 neurons from each area. We trained one CEBRA model per area and computed linear consistency between all pairs of embeddings. For intra-area comparison we sampled an additional 400 disjoint neurons. For each area we trained two CEBRA models on these two sets of neurons and computed their linear consistency. We repeated this process three times.

For comparisons across modalities (2P and NP) we sampled 400 neurons from each modality (which are disjoint, as above, because one set was sampled from 2P recordings and the other from NP recordings). We trained a multi-session CEBRA model with one encoder for 2P and one for NP in the same embedding space. For intra-area comparison we computed linear consistency between NP and 2P decoders from the same area. For interarea comparison we computed linear consistency between the NP encoder from one area and the 2P encoder from another and again considered all combinations of areas. We repeated this process three times.

For comparison of single- and multi-session training (Extended Data Fig. 7) we computed embeddings using encoder models with eight, 16, ..., 128 hidden units to vary the model size, and benchmarked eight, 16, ..., 128 latent dimensions. Hyperparameters, except for number of optimization steps, were selected according to either validation set decoding $R^2$ (rat) or accuracy (Allen). Consistency was reported as the point in training at which position decoding error was less than 7 cm for the first rat in the hippocampus dataset, and a decoding accuracy of 60% in the Allen dataset. For single-session training, four embeddings were trained independently on each individual animal whereas for multi-session training the embeddings

were trained jointly on all sessions. For multi-session training, the same number of samples was drawn from each session to learn an embedding invariant to the session ID. The consistency versus decoding error trade-off (Extended Data Fig. 7c) was reported as the average consistency across all 12 comparisons (Extended Data Fig. 7b) versus average decoding performance across all rats and data splits.

## Model comparisons

### pi-VAE parameter selection and modifications to pi-VAE

Because the original implementation of pi-VAE used a single time bin spiking rate as an input, we therefore modified their code to allow for larger time bin inputs and found that time window input with a receptive field of ten time bins (250 ms) gave higher consistency across subjects and better preserved the qualitative structure of the embedding (thereby outperforming the results presented by Zhou and Wei[5]; Extended Data Fig. 1). To do this we used the same encoder neural network architecture as that for CEBRA and modified the decoder to a 2D output (we call our modified version conv-pi-VAE). Note, we used this modified pi-VAE for all experiments except for the synthetic setting, for which there is no time dimension and thus the original implementation is sufficient.

The original implementation reported a median absolute error of 12 cm for rat 1 (the individual considered most in that work), and our implementation of time-windowed input with ten bins resulted in a median absolute error of 11 cm (Fig. 2). For hyperparameters we tested different epochs between 600 (the published value used) and 1,000, and learning rate between $1.0 \times 10^{-6}$ and $5.0 \times 10^{-4}$ via a grid search. We fixed hyperparameters as those that gave the highest consistency across subjects, which were training epochs of 1,000 and learning rate $2.5 \times 10^{-4}$. All other hyperparameters were retained as in the original implementation[5]. Note that the original paper demonstrated that pi-VAE is fairly robust across different hyperparameters. For decoding (Fig. 2) we considered both a simple kNN decoder (that we use for CEBRA) and the computationally more expensive Monte Carlo sampling method originally proposed for pi-VAE[5]. Our implementation of conv-pi-VAE can be found at https://github.com/AdaptiveMotorControlLab/CEBRA.

**autoLFADS parameter selection**

AutoLFADS[25] includes a hyperparameter selection and tuning protocol, which we used, and we also used the original implementation (https://github.com/snel-repo/autolfads-tf2/, https://github.com/neurallatents/nlb_tools/tree/main/examples/baselines/autolfads). For the rat hippocampus dataset we chopped the continuous spiking rate (25 ms bin size) into 250-ms-long segments with 225 ms overlap between segments to match the training setup for CEBRA, UMAP, *t*-SNE and pi-VAE. We used population-based training (PBT) for hyperparameter searches and constrained the search range to default values given in the original script (initial learning rate between $1.0 \times 10^{-5}$ and $5.0 \times 10^{-3}$, dropout rate 0–0.6, coordinated dropout rate 0.01–0.70, L2 generator weight between $1.0 \times 10^{-4}$ and 1.0, L2 controller weight between $1.0 \times 10^{-4}$ and 1.0, KL controller weight between $1.0 \times 10^{-6}$ and $1.0 \times 10^{-4}$ and KL initial condition weight between $1.0 \times 10^{-6}$ and $1.0 \times 10^{-3}$). The negative log-likelihood metric was used to select the best hyperparameters. Each generation of PBT consisted of 25 training epochs and we trained for a maximum of 5,000 epochs of batch size 100 while executing early stopping after awaiting 50 epochs. The PBT search was done using 20 parallel workers on each rat.

### UMAP parameter selection

For UMAP[11], following the parameter guide (umap-learn.readthedocs.io/), we focused on tuning the number of neighbours (*n_neighbors*) and minimum distance (*min_dist*). The *n_components* parameter was fixed to 2 and we used a cosine metric to make a fair comparison with CEBRA, which also used the cosine distance metric for learning. We performed a grid search with 100 total hyperparameter values in the range [2, 200] for *n_neighbors* and in the range [0.0001, 0.99] for *min_dist*. The highest consistency across runs in the rat hippocampus dataset was achieved with *min_dist* of 0.0001 and *n_neighbors* of 24. For the other datasets in Extended Data Fig. 3 we used the default value of *n_neighbors* as 15 and *min_dist* as 0.1.

### *t*-SNE parameter selection

For *t*-SNE[12] we used the implementation in openTSNE[58]. We performed a sweep on *perplexity* in the range [5, 50] and *early_exaggeration* in the range [12, 32] following the parameter guide, while fixing *n_components* as 2 and used a cosine metric for fair comparison with UMAP and CEBRA. We used PCA initialization to improve the run consistency of *t*-SNE[59]. The highest consistency across runs in the rat hippocampus dataset was achieved with *perplexity* of ten and *early_exaggeration* of 16.44. For the other datasets in Extended Data Fig. 3 we used the default value for *perplexity* of 30 and for *early_exaggeration* of 12.

## Decoding analysis

We primarily used a simple kNN algorithm, a nonparametric supervised learning method, as a decoding method for CEBRA. We used the implementation in scikit-learn[57]. We used a kNN regressor for continuous value regression and a kNN classifier for discrete label classification. For embeddings obtained with cosine metrics we used cosine distance metrics for kNN, and Euclidean distance metrics for those obtained in Euclidean space.

For the rat hippocampus data a kNN regressor, as implemented in scikit-learn[57], was used to decode position and a kNN classifier to decode direction. The number of neighbours was searched over the range [1, 4, 9, 16, 25] and we used the cosine distance metric. We used the $R^2$ score of predicted position and direction vector on the validation set as a metric to choose the best *n_neighbours* parameter. We report the median absolute error for positional decoding on the test set. For pi-VAE, we additionally evaluated decoding quality using the originally proposed decoding method based on Monte Carlo sampling, with the settings from the original article[5]. For autoLFADS, use of their default Ridge regression decoder[25] performed worse than our kNN decoder, which is why we reported all results for the kNN decoder. Note that UMAP, *t*-SNE and CEBRA-Time were trained using the full dataset without label information when learning the embedding, and we used the above split only for training and cross-validation of the decoder.

For direction decoding within the monkey dataset we used a Ridge classifier[57] as a baseline. The regularization hyperparameter was searched over $[10^{-6}, 10^{2}]$. For CEBRA we used a kNN classifier for decoding direction with $k$ searched over the range [1, 2500]. For conv-pi-VAE we searched for the best learning rate over $[1.0 \times 10^{-5}, 1.0 \times 10^{-3}]$. For position decoding we used Lasso[57] as a baseline. The regularization hyperparameter was searched over $[10^{-6}, 10^{2}]$. For conv-pi-VAE we used 600 epochs and searched for the best learning rates over $[5 \times 10^{-4}, 2.5 \times 10^{-4}, 0.125 \times 10^{-4}, 5 \times 10^{-5}]$ via a grid of $(x,y)$ space in 1 cm bins for each axis as the sampling process for decoding. For CEBRA we used kNN regression, and the number of neighbours $k$ was again searched over [1, 2500].

For the Allen Institute datasets we performed decoding (frame number or scene classification) for each frame from Video 1. Here we used a kNN classifier[57] with a population vector kNN as a baseline, similar to the decoding of orientation grating performed in ref. [46]. For CEBRA we used the same kNN classifier method as on CEBRA features. In both cases the number of neighbours, $k$, was searched over a range [1, 100] in an exponential fashion. We used neural data recorded during the first eight repeats as the train set, the ninth repeat for validation in choosing the

hyperparameter and the last repeat as the test set to report decoding accuracy. We also used a Gaussian naive Bayes decoder[57] to test linear decoding from the CEBRA model and neural population vector. Here we assumed uniform priors over frame number and searched over the range $[10^{-10}, 10^{3}]$ in an exponential manner for the *var_smoothing* hyperparameter.

For layer-specific decoding we used data from excitatory neurons in area VISp: layers 2/3 [Emx1-IRES-Cre, Slc17a7-IRES2-Cre]; layer 4 [Cux2-CreERT2, Rorb-IRES2-Cre, Scnn1a-Tg3-Cre]; and layers 5/6 [Nr5a1-Cre, Rbp4-Cre_KL100, Fezf2-CreER, Tlx3-Cre_PL56, Ntrsr1-cre].

### Neural Latents Benchmark

We tested CEBRA on the *mc-maze* 20 ms task from the Neural Latents Benchmark[37] (https://eval.ai/web/challenges/challenge-page/1256/leaderboard/3183). We trained the offset10-model with 48 output dimensions and [128, 256, 512] hidden units, as presented throughout the paper. We trained, in total, 48 models by additionally varying the temperature in [0.0001, 0.004] and time offsets from {1, 2}. We performed smoothing of input neural data using a Gaussian kernel with 50 ms s.d. Lastly, we took 45 embeddings from the trained models picked by the validation score, aligned the embeddings (using the Procrustes method[60]) and averaged them.

## Topological analysis

For the persistent cohomology analysis we utilized ripser.py[61]. For the hippocampus dataset we used 1,000 randomly sampled points from CEBRA-Behaviour trained with temperature 1, time offset 10 and minibatch size 512 for 10,000 training steps on the full dataset and then analysed up to 2D cohomology. Maximum distance considered for filtration was set to infinity. To determine the number of cocycles in each cohomology dimension with a significant lifespan we trained 500 CEBRA embeddings with shuffled labels, similar to the approach taken in ref. [33]. We took the maximum lifespan of each dimension across these 500 runs as a

threshold to determine robust Betti numbers, then surveyed the Betti numbers of CEBRA embeddings across three, eight, 16, 32 and 64 latent dimensions.

Next we used DREiMac[62] to obtain topology-preserving circular coordinates (radial angle) of the first cocycle ($H^1$) from the persistent cohomology analysis. Similar to above, we used 1,000 randomly sampled points from the CEBRA-Behaviour models of embedding dimensions 3, 8, 16, 32 and 64.

### Behaviour embeddings for video datasets

High-dimensional inputs, such as videos, need further preprocessing for effective use with CEBRA. First we used the recently presented DINO model[48] to embed video frames into a 768D feature space. Specifically we used the pretrained ViT/8 vision transformer model, which was trained by a self-supervised learning objective on the ImageNet database. This model is particularly well suited for video analysis and among the state-of-the-art models available for embedding natural images into a space appropriate for a kNN search[48], a desired property when making the dataset compatible with CEBRA. We obtained a normalized feature vector for each video frame, which was then used as the continuous behaviour variable for all further CEBRA experiments.

For scene labels, three individuals labelled each video frame using eight candidate descriptive labels allowing multilabel classes. We took the majority vote of these three individuals to determine the label of each frame. In the case of multilabels we considered this as a new class label. The above procedure resulted in ten classes of frame annotation.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Hippocampus dataset: https://crcns.org/data-sets/hc/hc-11/about-hc-11, and we used the preprocessing script from https://github.com/zhd96/pi-vae/blob/main/code/rat_preprocess_data.py. Primate dataset: https://gui.dandiarchive.org/#/dandiset/000127. Allen Institute dataset: Neuropixels data are available at https://allensdk.readthedocs.io/en/latest/visual_coding_neuropixels.html. The preprocessed 2P recordings are available at https://github.com/zivlab/visual_drift/tree/main/data.

# Code availability

Code: https://github.com/AdaptiveMotorControlLab/CEBRA. Code used to reproduce the figures: https://github.com/AdaptiveMotorControlLab/CEBRA-figures.

# References


1. Urai, A. E., Doiron, B., Leifer, A. M. & Churchland, A. K. Large-scale neural recordings call for new insights to link brain and behavior. *Nat. Neurosci.* **25**, 11–19 (2021).

2. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).

3. Jazayeri, M. & Ostojic, S. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr. Opin. Neurobiol.* **70**, 113–120 (2021).

4. Humphries, M. D. Strong and weak principles of neural dimension reduction. *Neuron. Behav. Data Anal. Theory* https://nbdt.scholasticahq.com/article/24619 (2020).

5. Zhou, D., & Wei, X. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. *Adv. Neural Inf. Process. Syst.*

https://proceedings.neurips.cc//paper/2020/file/510f2318f324cf07fce24c3a4b89c771-Paper.pdf (2020).

6. Vargas-Irwin, C. E. et al. Decoding complete reach and grasp actions from local primary motor cortex populations. *J. Neurosci.* **30**, 9659–9669 (2010).

7. Okorokova, E. V., Goodman, J. M., Hatsopoulos, N. G. & Bensmaia, S. J. Decoding hand kinematics from population responses in sensorimotor cortex during grasping. *J. Neural Eng*. **17**, 046035 (2020).

8. Yu, B. M. et al. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J. Neurophysiol.* **102**, 614–635 (2008).

9. Churchland, M. et al. Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).

10. Gallego, J. A. et al. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat. Commun.* **9**, 4233 (2018).

11. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. *J. Open Source Softw.* **3**, 861 (2018).

12. Maaten, L. V., Postma, E. O. & Herik, J. V. Dimensionality reduction: a comparative review. *J. Mach. Learn. Res.* **10**, 13 (2009).

13. Roeder, G., Metz, L. & Kingma, D. P. On linear identifiability of learned representations. *Proc. Mach. Learn. Res.* **139**, 9030–9039 (2021).

14. Hyvärinen, A., Sasaki, H. & Turner, R. E. Nonlinear ICA using auxiliary variables and generalized contrastive learning. *Proc. Mach. Learn. Res.* **89**, 859–868 (2019).

15. Sani, O. G., Abbaspourazad, H., Wong, Y. T., Pesaran, B. & Shanechi, M. M. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nat. Neurosci.* **24**, 140–149 (2020).

16. Klindt, D. A. et al. Towards nonlinear disentanglement in natural data with temporal sparse coding. *International Conference on Learning Representations* https://openreview.net/forum?id=EbIDjBynYJ8 (2021).

17. Pandarinath, C. et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2017).

18. Prince, L. Y., Bakhtiari, S., Gillon, C. J., & Richards, B. A. Parallel inference of hierarchical latent dynamics in two-photon calcium imaging of neuronal populations. Preprint at https://www.biorxiv.org/content/10.1101/2021.03.05.434105v1 (2021).

19. Gutmann, M. U. & Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.* **13**, 307–361 (2012).

20. Oord, A. V., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://doi.org/10.48550/arXiv.1807.03748 (2018).

21. Khosla, P. et al. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **33**, 18661–18673 (2020).

22. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. E. A simple framework for contrastive learning of visual representations. *Proc. Mach. Learn. Res.* **119**, 1597–1607 (2020).

23. Hälvä, H. et al. Disentangling identifiable features from noisy data with structured nonlinear ICA. *Adv. Neural Inf. Process. Syst.* **34**, 1624–1633 (2021).

24. Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M. & Brendel, W. Contrastive learning inverts the data generating process. *Proc. Mach. Learn. Res.* **139***,* 12979–12990 (2021).

25. Keshtkaran, M. R. et al. A large-scale neural network training framework for generalized estimation of single-trial population dynamics. *Nat. Methods* **19**, 1572–1577 (2022).

26. Grosmark, A. D. & Buzsáki, G. Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. *Science* **351**, 1440–1443 (2016).

27. Huxter, J. R., Burgess, N. & O'Keefe, J. Independent rate and temporal coding in hippocampal pyramidal cells. *Nature* **425**, 828–832 (2003).

28. Moser, E. I., Kropff, E. & Moser, M. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* **31**, 69–89 (2008).

29. Dombeck, D. A., Harvey, C. D., Tian, L., Looger, L. L. & Tank, D. W. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nat. Neurosci.* **13**, 1433–1440 (2010).

30. Curto, C. What can topology tell us about the neural code? *Bull. Am. Math. Soc* **54**, 63–78 (2016).

31. Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A. & Fiete, I. R. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nat. Neurosci.* **22**, 1512–1520 (2019).

32. Silva, V. D., Morozov, D. & Vejdemo-Johansson, M. Persistent cohomology and circular coordinates. *Discrete Comput. Geom.* **45**, 737–759 (2009).

33. Gardner, R. J. et al. Toroidal topology of population activity in grid cells. *Nature* **602**, 123–128 (2022).

34. Chowdhury, R. H., Glaser, J. I. & Miller, L. E. Area 2 of primary somatosensory cortex encodes kinematics of the whole arm. *eLife* **9**, e48198 (2019).

35. Prud'homme, M. J. & Kalaska, J. F. Proprioceptive activity in primate primary somatosensory cortex during active arm reaching movements. *J. Neurophysiol.* **72**, 2280–2301 (1994).

36. London, B. M. & Miller, L. E. Responses of somatosensory area 2 neurons to actively and passively generated limb movements. *J. Neurophysiol.* **109**, 1505–1513 (2013).

37. Pei, F. et al. Neural Latents Benchmark '21: Evaluating latent variable models of neural population activity. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* https://openreview.net/forum?id=KVMS3fl4Rsv (2021).

38. Berens, P. et al. Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS Comput. Biol.* **14**, e1006157 (2018).

39. Hafting, T., Fyhn, M., Molden, S., Moser, M. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).

40. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).

41. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2011).

42. Menegas, W. et al. Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *eLife* **4**, e10032 (2015).

43. Hubel, D. H. & Wiesel, T. N. Ferrier lecture – functional architecture of macaque monkey visual cortex. *Proc. R. Soc. Lond. B Biol. Sci.* **198**,

1–59 (1977).

44. Niell, C. M., Stryker, M. P. & Keck, W. M. Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* **28**, 7520–7536 (2008).

45. Ringach, D. L. et al. Spatial clustering of tuning in mouse primary visual cortex. *Nat. Commun.* **7**, 12270 (2016).

46. de Vries, S. E. et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nat. Neurosci.* **23**, 138–151 (2019).

47. Siegle, J. H. et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).

48. Caron, M. et al. Emerging properties in self-supervised vision transformers. *IEEE/CVF International Conference on Computer Vision* 9630–9640 (2021).

49. Esfahany, K., Siergiej, I., Zhao, Y. & Park, I. M. Organization of neural population code in mouse visual system. *eNeuro* **5**, ENEURO.0414-17.2018 (2018).

50. Jin, M. & Glickfeld, L. L. Mouse higher visual areas provide both distributed and specialized contributions to visually guided behaviors. *Curr. Biol.* **30**, 4682–4692 (2020).

51. Dinh, L., Sohl-Dickstein, J. N. & Bengio, S. Density estimation using Real NVP. *International Conference on Learning Representations* https://openreview.net/pdf?id=HkpbnH9lx (2017).

52. Deitch, D., Rubin, A. & Ziv, Y. Representational drift in the mouse visual cortex. *Curr. Biol.* **31**, 4327–4339 (2021).

53. Wang, T. & Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *Proc. Mach. Learn. Res.* **119**, 9929–9939 (2020).

54. Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs). Preprint at https://doi.org/10.48550/arXiv.1606.08415 (2016).

55. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Neural Inf. Process. Syst*. **32**, 8024–8035 (2019).

56. Walt, S. V., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

57. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

58. Policar, P. G., Stražar, M. & Zupan, B. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. Preprint at *bioRxiv* https://doi.org/10.1101/731877 (2019).

59. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both *t*-SNE and UMAP. *Nat. Biotechnol.* **39**, 156–157 (2021).

60. Schönemann, P. H. A generalized solution of the orthogonal procrustes problem. *Psychometrika* **31**, 1–10 (1966).

61. Tralie, C. J., Saul, N. & Bar-On, R. Ripser.py: a lean persistent homology library for Python. *J. Open Source Softw.* **3**, 925 (2018).

62. Tralie, C. J., Mease, T. & Perea, J. Dreimac: dimension reduction with Eilenberg–Maclane coordinates. GitHub https://github.com/ctralie/DREiMac/tree/cdd6d02ba53c3597a931db9da478fd198d6ed00f (2018).

# Acknowledgements

Funding was provided by SNSF grant no. 310030_201057, The Vallee Foundation and a Novartis Foundation for Medical-Biological Research Young Investigator Grant to M.W.M., a Google PhD Fellowship to S.S. and the German Academic Exchange Service to J.H.L. S.S. acknowledges the

IMPRS-IS Tübingen and ELLIS PhD programme, and J.H.L. thanks the TUM Programme in Neuroengineering. M.W.M. is the Bertarelli Foundation Chair of Integrative Neuroscience. We thank L. Eck for help with theory; A. Mathis for advice on topological data analysis; M. Bethge, R. Zimmerman, D. Paiton and W. Brendel for general discussions; C. Benquet, R. González and J. Lauer for feedback on code; G. Kane for advice on statistical analysis; and M. Frey, D. Kobak and J. Macke for comments on the manuscript. The funders had no role in the conceptualization, design, data collection, analysis, decision to publish or preparation of the manuscript.

# Funding

Open access funding provided by EPFL Lausanne.

# Author information

Author notes

1. These authors contributed equally: Steffen Schneider, Jin Hwa Lee

## Authors and Affiliations

1. Brain Mind Institute & Neuro X Institute, École Polytechnique Fédérale de Lausanne, Geneva, Switzerland

   Steffen Schneider, Jin Hwa Lee & Mackenzie Weygandt Mathis

## Contributions

M.W.M. and S.S. conceptualized the project. S.S., J.H.L. and M.W.M. were responsible for the methodology. S.S., J.H.L. and M.W.M. were responsible for the software. S.S. was responsible for the theory. S.S. and J.H.L. undertook the formal analysis. S.S. and J.H.L. did the investigation. J.H.L., S.S. and M.W.M. were responsible for data curation. M.W.M. wrote the

original draft. M.W.M., S.S. and J.H.L. wrote and edited the final version of the article.


### Corresponding author

Correspondence to Mackenzie Weygandt Mathis.


## Ethics declarations

### Competing interests

S.S. and M.W.M. have filed a patent pertaining to the method presented in this work (filing no. 63/302,670). The remaining author declares no competing interests.

## Peer review

### Peer review information


*Nature* thanks Anne Churchland, Benjamin Dunn and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.


## Additional information


**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


## Extended data figures and tables

### Extended Data Fig. 1 Overview of datasets, synthetic data, & original pi-VAE implementation vs. modified conv-pi-VAE.

**a**, We generated synthetic datasets similar to Fig. 1b with additional variations in the noise distributions in the generative process. We benchmarked the reconstruction score of the true latent using CEBRA and pi-VAE ($n = 100$ seeds) on the generated synthetic datasets. CEBRA showed higher and less variable reconstruction scores than pi-VAE in all noise types (one-sided Welch's t-test, corrected using the Holm-Bonferroni method, t and p-values indicated on the plot). **b**, Example visualization of the reconstructed latents from CEBRA and pi-VAE on different synthetic dataset types. **c**, We benchmarked and demonstrate the abilities of CEBRA on four datasets. Rat-based electrophysiology data[26], where the animal transverse a 1.6 meter linear track "leftwards" or "rightwards". Two mouse-based datasets: one 2-photon calcium imaging passively viewing dataset[46], and one with the same stimulus but recorded with Neuropixels[47]. A monkey-based electrophysiology dataset of centre out reaching from Chowdhury et al.[34], and processed to trial data as in ref. [52]. **d**, Conv-pi-VAE showed improved performance, both with labels (Wilcoxon signed-rank test, $P = 0.0341$) and without labels Wilcoxon signed-rank test, $P = 0.0005$). Example runs/embeddings the consistency across rats, with **e**, consistency across rats, from target to source, as computed in Fig. 1. Cartoon animals are adapted from scidraw.io.

## Extended Data Fig. 2 Hyperparameter changes on visualization and consistency.

**a**, Temperature has the largest effect on visualization (vs. consistency) of the embedding as shown by a range from 0.1 to 3.21 (highest consistency for rat 1), as can be appreciated in 3D (top) and post FastICA into a 2D embedding (middle). Bottom row shows the corresponding change on mean consistency, and in **b**, the variance can be noted. Orange line denotes the median and black dots are individual runs (subject consistency: 10 runs with 3 comparisons per rat; run consistency: 10 runs, each compared to 9 remaining runs).

## Extended Data Fig. 3 CEBRA produced consistent, highly decodable embeddings.

**a**, Additional rat data shown for all algorithms we benchmarked (see Methods). For CEBRA-Behaviour, we used temperature 1, time offset 10, batch size 512 and 10k training steps. For CEBRA-Time, we used temperature 2.25, time offset 10, batch size 512 and 4k training steps. For UMAP, we used the cosine metric and *min_dist* of 0.99 and *n_neighbors* of 31. For t-SNE we used cosine metric and *perplexity* of 29. For conv-pi-VAE, we trained 1000 epochs with learning rate 2.5 x $10^{-4}$. For autoLFADS we used the in-built ray-tune framework for finding optimal hyperparameters. CEBRA was trained with output latent 3D (the minimum) and all other methods were trained with a 2D latent.

## Extended Data Fig. 4 Additional metrics used for benchmarking consistency.

Comparisons of all algorithms along different metrics for rats 1, 2, 3, 4. The orange line is median across n = 10 runs, black circles denote individual runs. Each run is the average over three non-overlapping test splits.

## Extended Data Fig. 5 Hypothesis testing with CEBRA.

**a**, Example data from a hippocampus recording session (rat 1). We tested possible relationships between three experimental variables (rat location, velocity, movement direction) and the neural recordings (120 neurons, not shown). **b**, Relationship between velocity and position. **c**, We trained CEBRA with three-dimensional outputs on every single experimental variable (main diagonal) and every combination of two variables. All variables are treated as 'continuous' in this experiment. We compared original to shuffled variables (shuffling is done by permuting all samples over the time dimension) as a control. We projected the original three-dimensional space onto the first principal components. We show the minimum value of the InfoNCE loss on the trained embedding for all combinations in the confusion matrix (lower number is better). Either velocity or direction, paired with position information is needed for maximum structure in the embedding (highlighted, coloured), yielding lowest InfoNCE error. **d**, Using an eight-dimensional CEBRA embedding did not qualitatively alter the results. We again report the first two principal components as well as InfoNCE training error upon convergence, and find

non-trivial embeddings with lowest training error for combinations of direction/velocity and position. **e**, The InfoNCE metric can serve as the goodness of fit metric, both for hypothesis testing and identifying decodable embeddings. We trained CEBRA in discovery-driven mode with 32 latent dimensions. We compared the InfoNCE loss (left, middle) between various hypotheses. Low InfoNCE was correlated with low decoding error (right).

## Extended Data Fig. 6 Persistence across dimensions.

**a**, For each dimension of CEBRA-Behaviour embedding from the rat hippocampus dataset Betti numbers were computed by applying persistent cohomology. The coloured dots are lifespans observed in hypothesis based CEBRA-Behaviour. To rule out noisy lifespans, we set a threshold (coloured diagonal lines) as maximum lifespan based on 500 seeds of shuffled-CEBRA embedding for each dimension. **b**, The topology preserving circular coordinates using the first co-cycle from persistent cohomology analysis on the CEBRA embedding of each dimension is shown (see Methods). The colours indicate position and direction of the rat at the corresponding CEBRA embedding points. **c**, The radial angle of each embedding point obtained from **b** and the corresponding position and direction of the rat.

## Extended Data Fig. 7 Multi-session training and rapid decoding.

**a**, Top: hippocampus dataset, single animal vs. multi-animal training shows an increase in consistency across animals. Bottom: same for Allen dataset, 4 mice. **b**, Consistency matrix single vs. multi-session training for hippocampus (32D embedding) and Allen data (128D embedding) respectively. Consistency is reported at the point in training where the average position decoding error is less than 14 cm (corresponds to 7 cm error for rat 1), and a decoding accuracy of 60% on the Allen dataset. **c**, Comparison of decoding metrics for single or multi-session training at various consistency levels (averaged across all 12 comparisons). Models were trained for 5,000 (single) or 10,000 (multi-session) steps with a 0.003 learning rate; batch size was 7,200 samples per session. Multi-session training requires longer training or higher learning rates to obtain the same

accuracy due to the 4-fold larger batch size, but converges to same decoding accuracy. We plot points at intervals of 500 steps ($n = 10$ seeds); training progresses from lower right to upper left corner within both plots. **d**, We demonstrate that we could also adapt to an unseen dataset; here, 3 rats were used for pretraining, and rat 4 was used as a held-out test. The grey lines indicate models trained from scratch (random initialization). We also tested fine-tuning only the input embedding (first layer) or the full model, as the diagram, left, describes. We measured the average time (mean ± s.d.) to adapt 100 steps (0.65 ± 0.13 s) and 500 steps (3.07 ± 0.61 s) on 40 repeated experiments.

## Extended Data Fig. 8 Somatosensory cortex decoding from primate recordings.

**a**, We compare CEBRA-Behaviour with the cosine similarity and embeddings on the sphere reproduced from Fig. 3b (left) against CEBRA-Behaviour trained with the MSE loss and unnormalized embeddings. The embeddings of trials ($n = 364$) of each direction were post-hoc averaged. **b**, CEBRA-Behaviour trained with *x,y* position of the hand. Left panel is colour- coded to changes in *x* position and right panel is color-coded to changes in *y* position. **c**, CEBRA-Time without any external auxiliary variables. As in **b**, left and right are colour-coded to *x* and *y* position, respectively. **d**, Decoding performance of target direction using CEBRA-Behaviour, conv-pi-VAE and a linear classifier. CEBRA-Behaviour shows significantly higher decoding performance than the linear classifier (one-way ANOVA, $F(2,75) = 3.37$, $P < 0.05$ with Post Hoc Tukey significant difference $P < 0.05$). **e**, Loss (InfoNCE) vs. training iteration for CEBRA-Behaviour with position, direction, active or passive, and position+direction labels (and shuffled labels) for all trials (left) or only active trials (right), or active trials with a MSE loss. **f**, Additional decoding performance results on position and direction-trained CEBRA models with all trial types. For each case, we trained and evaluated 5 seeds represented by black dots and the orange line represents the median. **g**, Results on the mc-maze 20 ms benchmark.

## Extended Data Fig. 9 CEBRA produces consistent, highly decodable embeddings.

**a**, Additional 4 sessions with the most neurons in the Allen visual dataset calcium recording shown for all algorithms we benchmarked (see Methods). For CEBRA-Behaviour and CEBRA-Time, we used 3D, temperature 1, time offset 10, batch size 512 and 10k training steps. For UMAP, we used a cosine metric and *n_neighbors* 15 and *min_dist* 0.1. For t-SNE, we used a cosine metric and *perplexity 30*. For conv-pi-VAE, we trained with 600 epochs, a batch size of 200 and a learning rate 5 x $10^{-4}$. autoLFADS was trained with ray-tune parameter selection and the resulting factors were transformed with PCA to generate the visualization. All methods used 10-time-bins input. CEBRA was trained with 3D latent and all other methods were obtained with an equivalent 2D latent dimension. To align for visualization, we aligned to mouse 1, except for conv-pi-VAE without labels and for autoLFADS, which visually looked best when aligned to mouse 4.

## Extended Data Fig. 10 Spikes and calcium signalling reveal similar embeddings.

**a**, Consistency between the single and jointly trained embeddings. **b**, Consistency of embeddings from two recording modalities, when a single modality was trained independently and/or jointly trained. CEBRA can find 'common latents' even without joint training. Data is also presented in Fig. 4e, h, but here plotted together to show improvement with joint training; for **a** and **b**, for each neuron number we have $n = 5$ shuffles, mean ± s.e.m. **c-f**, Consistency across modalities and areas for CEBRA-Behaviour and -Time (as computed in Fig. 4i–k). The purple dots indicate mean of intra-V1 scores and inter-V1 scores (inter-V1 vs intra-V1 one-sided Welch's t-test; 2P (Behaviour): $t(10.6) = 1.52$, $P = 0.081$, 2P (Time): $t(44.3) = 4.26$, $P = 0.0005$, NP (Behaviour): $t(11.6) = 2.83$, $P = 0.0085$, NP (Time): $t(8.9) = 15.51$, $P < 0.00001$) **g**, CEBRA + kNN decoding performance (see Methods) of CEBRA embeddings of different output embedding dimensions, from calcium (2P) data or Neuropixels (NP), as denoted; for each neuron number we have $n = 5$ shuffles, mean ± s.e.m. **h**, Decoding

accuracy measured by considering predicted frame being within 1 s difference to true frame using CEBRA (2P only), jointly trained (2P+NP), or a baseline population vector kNN decoder (using time window 33 ms (single frame), or 330 ms (10 frame receptive field)); for each neuron number we have $n = 5$ shuffles, mean ± s.e.m. (**i**): Single frame performance and quantification using CEBRA 1 frame receptive field (NP data), or baseline models, $n = 900$ video frames. **j**, CEBRA-Behaviour used the DINO features as auxiliary labels and DINO-shuffled used the shuffled DINO features. We shuffled the frame order of DINO features within a repeat. Same shuffled order was use for all repeats. Colour code is frame number from the movie. The prediction is considered as true if the predicted frame is within 1 s from the true frame, and the accuracy (%) is noted next to the embedding. For mouse ID 1–4: 337, 353, 397, 475 neurons were recorded, respectively. **k**, Decoding performance from 2P data from different visual cortical areas from different layers using a 10-frame-window, 128D CEBRA-Behaviour model using DINO features; for each neuron number we have $n = 5$ shuffles, mean± s.e.m.

## Supplementary information

### Supplementary Information

Attached in a single PDF are Supplementary Notes 1 and 2, which provide extended discussions on identifiability and the theoretical guarantees of CEBRA, respectively, and Tables 1–6, which provide statistical support to the conclusions drawn in the main manuscript.

### Reporting Summary

### Supplementary Video 1

Corresponding to Fig. 2d. CEBRA-Behaviour trained with position and direction on rat 1. Video is in 2× real time.

### Supplementary Video 2

Corresponding to Fig. 5b. The left-hand panels show example calcium traces from two-photon imaging (top) and spikes from Neuropixels recording (bottom) of primary visual cortex while the video was shown to mice (here, we randomly picked neurons to visualize the pseudomouse). The centre panel shows an embedding space constructed by jointly training a CEBRA-Behaviour model with two-photon and Neuropixels recordings using DINO frame features as labels. The trace is an embedding of a held-out test repeat from the Neuropixels recording. The colour map indicates the frame number of the 30-s-long video (30 Hz). The final panels show true video (top) and predicted frame sequence (bottom) using a kNN decoder on CEBRA-Behaviour embedding from the test set. Video is in real time.

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

## About this article

### Cite this article

Schneider, S., Lee, J.H. & Mathis, M.W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023).

https://doi.org/10.1038/s41586-023-06031-6

- Received: 30 March 2022
- Accepted: 28 March 2023
- Published: 03 May 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-06031-6

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

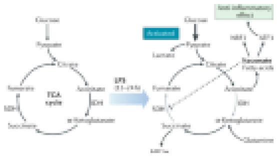
Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

## Brain dynamics uncovered using a machine-learning algorithm

Nature Research Briefing 03 May 2023

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-06031-6



## Itaconate: the poster child of metabolic reprogramming in macrophage function

31 January 2019

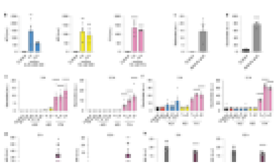
Luke A. J. O’Neill & Maxim N. Artyomov



## IFNγ: signalling, epigenetics and roles in immunity, metabolism, disease and cancer immunotherapy

19 June 2018

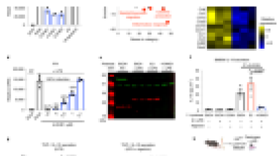
Lionel B. Ivashkiv



## Comparative evaluation of itaconate and its derivatives reveals divergent inflammasome and type I interferon regulation in macrophages

25 May 2020

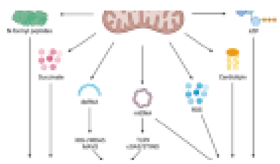
Amanda Swain, Monika Bambouskova, … Maxim N. Artyomov



[Mitohormesis reprogrammes macrophage metabolism to enforce tolerance]

20 May 2021

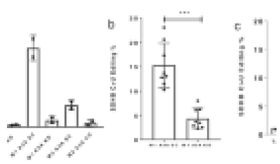
Greg A. Timblin, Kevin M. Tharp, … Kaoru Saijo



[Mitochondria and cell death-associated inflammation]

29 November 2022

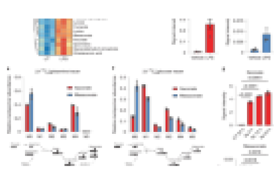
Esmee Vringer & Stephen W. G. Tait



[RNA editing enzyme APOBEC3A promotes pro-inflammatory M1 macrophage polarization]

22 January 2021

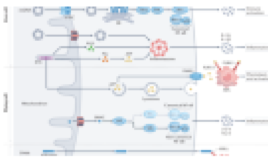
Emad Y. Alqassim, Shraddha Sharma, … Bora E. Baysal



[Mesaconate is synthesized from itaconate and exerts immunomodulatory effects in macrophages](#)

02 June 2022

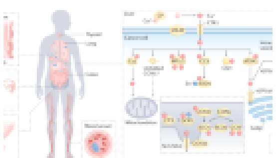
Wei He, Antonia Henne, … Karsten Hiller



[Mitochondrial control of inflammation](#)

25 July 2022

Saverio Marchi, Emma Guilbaud, … Lorenzo Galluzzi



[Connecting copper and cancer: from transition metal signalling to metalloplasia](#)

11 November 2021

Eva J. Ge, Ashley I. Bush, … Christopher J. Chang


- Article
- Open Access
- Published: 26 April 2023


# A druggable copper-signalling pathway that drives inflammation

- Stéphanie Solier,
- Sebastian Müller,
- Tatiana Cañeque,
- Antoine Versini,
- Arnaud Mansart,
- Fabien Sindikubwabo,
- Leeroy Baron,
- Laila Emam,
- Pierre Gestraud,
- G. Dan Pantoș,
- Vincent Gandon,
- Christine Gaillet,
- Ting-Di Wu,
- Florent Dingli,
- Damarys Loew,
- Sylvain Baulande,
- Sylvère Durand,
- Valentin Sencio,
- Cyril Robil,
- François Trottein,
- David Péricat,
- Emmanuelle Näser,
- Céline Cougoule,
- Etienne Meunier,
- Anne-Laure Bègue,
- Hélène Salmon,
- Nicolas Manel,
- Alain Puisieux,
- Sarah Watson,
- Mark A. Dawson,
- Nicolas Servant,
- Guido Kroemer,
- Djillali Annane &
- …
- Raphaël Rodriguez


Show authors

*Nature* **volume 617**, pages 386–394 (2023)

- 17k Accesses
- 1 Citations
- 154 Altmetric
- Metrics details


# Abstract

Inflammation is a complex physiological process triggered in response to harmful stimuli[1]. It involves cells of the immune system capable of clearing sources of injury and damaged tissues. Excessive inflammation can occur as a result of infection and is a hallmark of several diseases[2,3,4]. The molecular bases underlying inflammatory responses are not fully understood. Here we show that the cell surface glycoprotein CD44, which marks the acquisition of distinct cell phenotypes in the context of development, immunity and cancer progression, mediates the uptake of metals including copper. We identify a pool of chemically reactive copper(ii) in mitochondria of inflammatory macrophages that catalyses NAD(H) redox cycling by activating hydrogen peroxide. Maintenance of $NAD^+$ enables metabolic and epigenetic programming towards the inflammatory state. Targeting mitochondrial copper(ii) with supformin (LCC-12), a rationally designed dimer of metformin, induces a reduction of the NAD(H) pool, leading to metabolic and epigenetic states that oppose macrophage activation. LCC-12 interferes with cell plasticity in other settings and reduces inflammation in mouse models of bacterial and viral infections. Our work highlights the central role of copper as a regulator of cell plasticity and unveils a therapeutic strategy based on metabolic reprogramming and the control of epigenetic cell states.


# Main

Inflammation is a complex physiological process that enables clearance of pathogens and repair of damaged tissues. However, uncontrolled

inflammation driven by macrophages and other immune cells can result in tissue injury and organ failure. Effective drugs against severe forms of inflammation are scarce[5,6], and there is a need for therapeutic innovation[7].

The plasma membrane glycoprotein CD44 is the main cell surface receptor of hyaluronates[8,9,10]. It has been associated with biological programmes[11] that involve cells capable of acquiring distinct phenotypes independently of genetic alterations, which is commonly defined as cell plasticity[12,13]. For instance, inflammatory macrophages are marked by increased expression of CD44 and its functional implication in this context has been demonstrated[14,15]. However, the mechanisms by which CD44 and hyaluronates influence cell biology remain elusive[14,16,17,18]. The recent discovery that CD44 mediates the endocytosis of iron-bound hyaluronates in cancer cells links membrane biology to the epigenetic regulation of cell plasticity, where increased iron uptake promotes the activity of α-ketoglutarate (αKG)-dependent demethylases involved in the regulation of gene expression[19]. Hyaluronates have been shown to induce the expression of pro-inflammatory cytokines in alveolar macrophages (AMs)[20], and macrophage activation relies on complex regulatory mechanisms occurring at the chromatin level[21,22,23]. This body of work raises the question of whether a general mechanism involving CD44-mediated metal uptake regulates macrophage plasticity and inflammation.

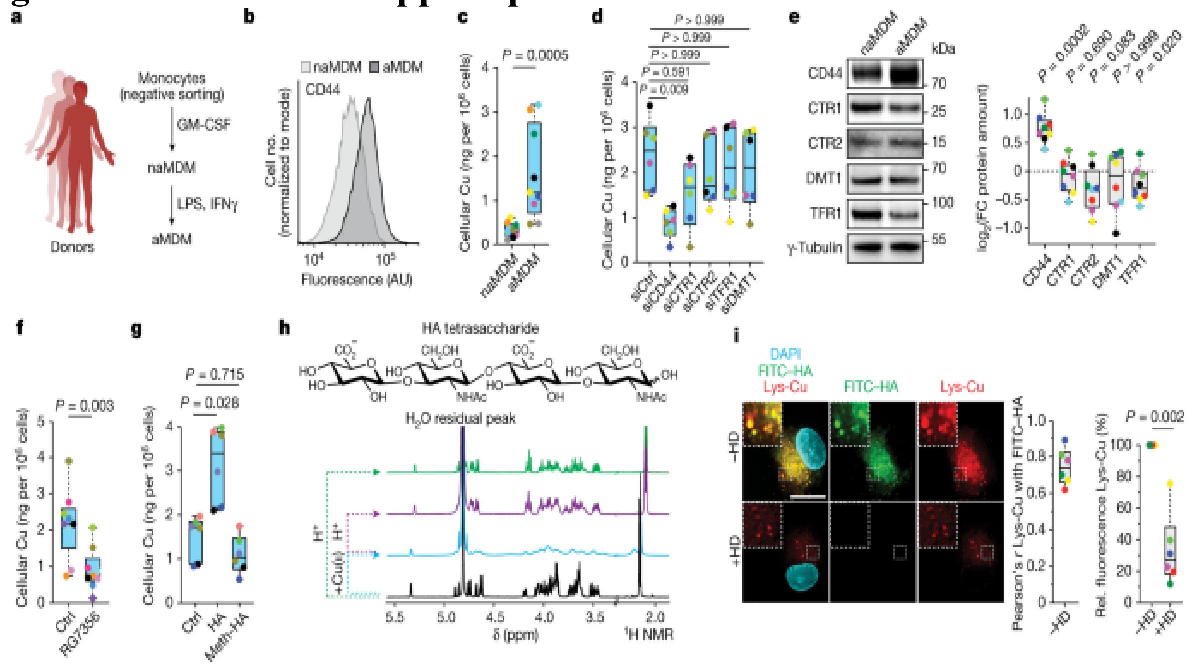
Here we show that macrophage activation is characterized by an increase of mitochondrial copper(ii), which occurs as a result of CD44 upregulation. Mitochondrial copper(ii) catalyses NAD(H) redox cycling, thereby promoting metabolic changes and ensuing epigenetic alterations that lead to an inflammatory state. We developed a metformin dimer that inactivates mitochondrial copper(ii). This drug induces metabolic and epigenetic shifts that oppose macrophage activation and dampen inflammation in vivo.

# CD44 mediates cellular uptake of copper

To study the role of metals in immune cell activation, we generated inflammatory macrophages using human primary monocytes isolated from blood (Fig. 1a). Activated monocyte-derived macrophages (aMDMs) were

characterized by the upregulation of CD44, CD86 and CD80, together with a distinct cell morphology (Fig. 1b and Extended Data Fig. 1a–c).

**Fig. 1: CD44 mediates copper uptake.**




**a**, Experimental setup used to generate inflammatory monocyte-derived macrophages (MDMs). **b**, Flow cytometry of CD44 in MDMs. Data are representative of $n = 13$ donors. AU, arbitrary units. **c**, ICP-MS of cellular copper in MDMs ($n = 9$ donors). **d**, ICP-MS of cellular copper in aMDMs with short interfering RNA (siRNA) knockdown of indicated receptors and transporters ($n = 6$ donors). Copper transporter 1 (CTR1) is encoded by *SLC31A1*, CTR2 is encoded by *SLC31A2*, transferrin receptor 1 (TFR1) is encoded by *TFRC*, and divalent metal transporter 1 (DMT1) is encoded by *SLC11A2*. siCtrl, control siRNA. **e**, Representative western blots of metal transporters in MDMs ($n = 7$ donors). FC, fold change. **f**, ICP-MS of cellular copper in MDMs treated with anti-CD44 antibody RG7356 during activation ($n = 7$ donors). **g**, ICP-MS of cellular copper in MDMs treated with hyaluronate (0.6–1 MDa) (HA) or permethylated hyaluronate (meth–HA) during activation ($n = 6$ donors). **h**, Molecular structure of hyaluronate tetrasaccharide (top) and $^1H$ NMR spectra (bottom) of copper–hyaluronate complexation experiment, recorded at 310 K in $D_2O$. **i**, Fluorescence microscopy of a lysosomal copper(ii) probe (Lys-Cu) and FITC–hyaluronate in aMDMs treated with hyaluronidase (HD). At least 30 cells were

quantified per donor ($n = 6$ donors). Scale bar, 10 µm. Rel., relative. **c**,**e**,**f**,**i**, Two-sided Mann–Whitney test. **d**,**g**, Kruskal–Wallis test with Dunn's post test. In all box plots in the main figures, boxes represent the interquartile range, centre lines represent medians and whiskers indicate the minimum and maximum values. In graphs, each coloured dot represents an individual donor for a given panel.

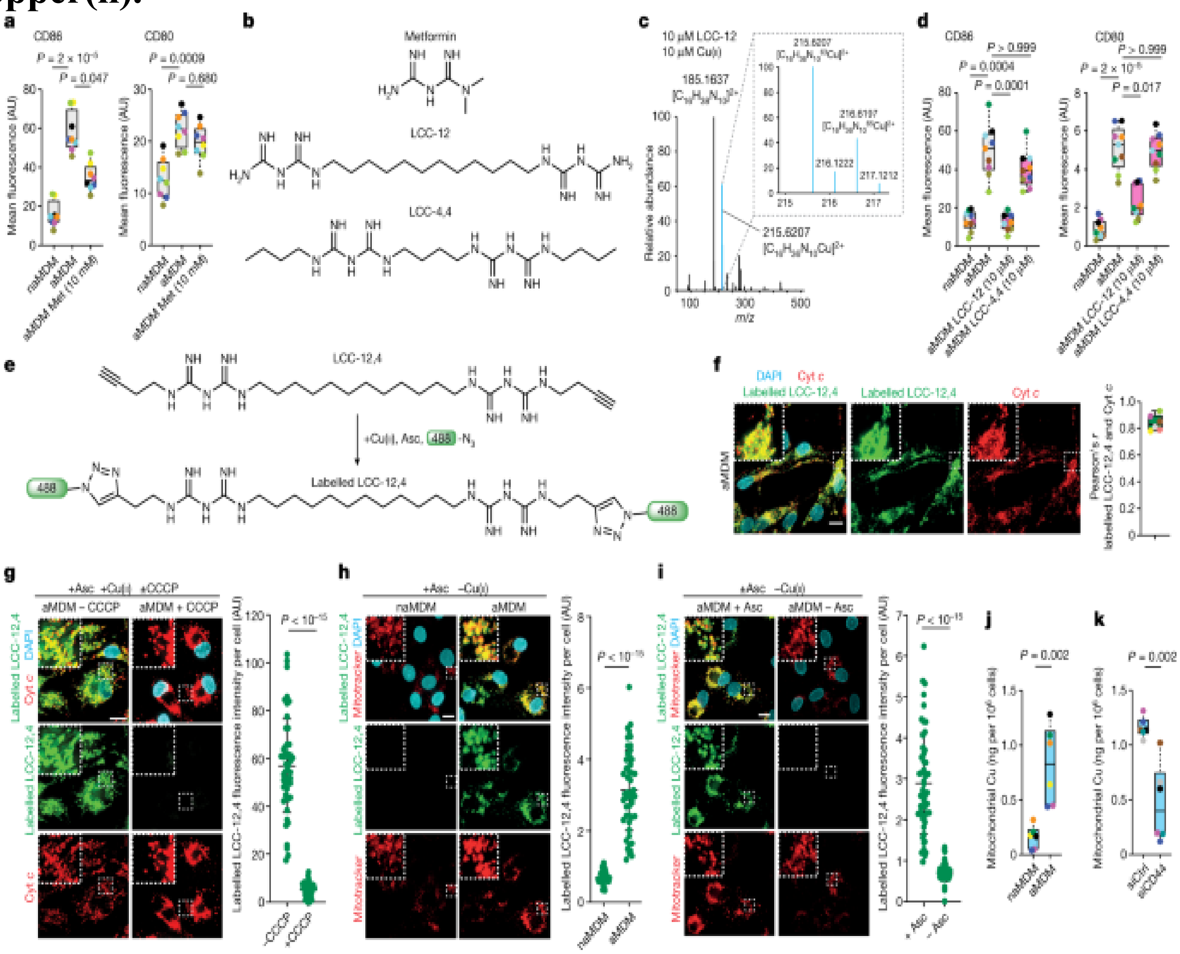
Source data

Using inductively coupled plasma mass spectrometry (ICP-MS), we detected higher levels of cellular copper, iron, manganese and calcium in aMDMs compared with non-activated MDMs (naMDMs) (Fig. 1c and Extended Data Fig. 1d). In contrast to other metal transporters, knocking down CD44 antagonized metal uptake (Fig. 1d and Extended Data Fig. 1e,f) and, unlike CD44, levels of these other metal transporters did not increase upon macrophage activation (Fig. 1e). Of note, levels of these transporters remained unchanged under CD44-knockdown conditions (Extended Data Fig. 1g). Treating MDMs with an anti-CD44 antibody[24] antagonized metal uptake upon activation (Fig. 1f and Extended Data Fig. 2a). Conversely, supplementing cells with hyaluronate upon activation increased metal uptake, whereas addition of a permethylated hyaluronate[25], which is less prone to metal binding, had no effect (Fig. 1g and Extended Data Fig. 2b). Inflammatory macrophages were also characterized by the upregulation of hyaluronate synthases (HAS) and the downregulation of the copper export proteins ATP7A and ATP7B (Extended Data Fig. 2c). Nuclear magnetic resonance revealed that hyaluronate interacts with copper(ii) and that this interaction can be reversed by lowering the pH (Fig. 1h). Fluorescence microscopy showed that labelled hyaluronate colocalized with a lysosomal copper(ii) probe[26] in aMDMs (Fig. 1i). Cotreatment with hyaluronidase—which degrades hyaluronates—or knocking down CD44 reduced lysosomal copper(ii) staining (Fig. 1i and Extended Data Fig. 2d). In aMDMs, the copper transporter CTR2 colocalized with the endolysosomal marker LAMP2, and CTR2 knockdown led to increased lysosomal copper(ii) staining (Extended Data Fig. 2e,f). Collectively, these data indicate that in aMDMs, CD44 mediates the endocytosis of specific metals bound to hyaluronate, including copper.

## Mitochondrial Cu(ii) regulates cell plasticity

We evaluated the capacity of copper(i) and copper(ii) chelators, including ammonium tetrathiomolybdate (ATTM), d-penicillamine (D-Pen), EDTA and trientine to interfere with macrophage activation. We also studied metformin, a biguanide used for the treatment of type-2 diabetes, because it can form a bimolecular complex[27] with copper(ii). Metformin partially antagonized CD86 upregulation, albeit at high concentrations, in contrast to the marginal effects of other copper-targeting molecules (Fig. 2a and Extended Data Fig. 3a).

**Fig. 2: Development of a small molecule inactivator of mitochondrial copper(ii).**




**a**, Flow cytometry of cell surface markers in MDMs treated with metformin (Met) ($n = 8$ donors). **b**, Molecular structures of metformin, LCC-12 and LCC-4,4. **c**, HRMS of a Cu–LCC-12 complex. **d**, Flow cytometry of cell

surface markers in MDMs treated with LCC-12 or LCC-4,4 ($n = 9$ donors). **e**, Experimental procedure of in-cell labelling of LCC-12,4. **f**, Fluorescence microscopy of labelled LCC-12,4 in aMDMs ($n = 6$ donors). At least 50 cells were quantified per donor. Cyt *c*, cytochrome *c*. **g**, Fluorescence microscopy of labelled LCC-12,4 in aMDMs treated with CCCP. **h**, Fluorescence microscopy of labelled LCC-12,4 in MDMs. In-cell labelling is performed with ascorbate and without added copper(ii). **i**, Fluorescence microscopy of labelled LCC-12,4 in aMDMs. In-cell labelling is performed in the presence or absence of ascorbate (asc) and without added copper(ii). **j**, ICP-MS of mitochondrial copper in MDMs ($n = 6$ donors). **k**, ICP-MS of mitochondrial copper in aMDMs under CD44-knockdown conditions ($n = 6$ donors). **g**–**i**, Two-sided unpaired *t*-test, representative of $n = 3$ donors. Data are mean ± s.d. **a**,**d**, Kruskal–Wallis test with Dunn's post test. **j**,**k**, Two-sided Mann–Whitney test. In graphs, each coloured dot represents an individual donor for a given panel. Scale bars, 10 μm.

Source data

To reduce the entropic cost inherent to the formation of bimolecular $Cu(Met)_2$ complexes[27,28], we tethered two biguanides with methylene-containing linkers to produce the lipophilic copper clamps LCC-12 and LCC-4,4 (Fig. 2b), which contain 12 and 4 linking methylene groups, respectively. LCC-4,4 displays distal butyl substituents to exhibit a lipophilicity similar to that of LCC-12. We compared simulated structures of copper(ii) complexes with the lowest energies using molecular dynamics and discrete Fourier transform with a $Cu(Met)_2$ complex using the crystal structure of the latter as benchmark[28] (Extended Data Fig. 3b). Cu–LCC-12 adopted a geometry similar to that of $Cu(Met)_2$, whereas Cu–LCC-4,4 lacked bonding angle symmetry and exhibited imine–copper bonds out of plane. The calculated free energy of Cu–LCC-4,4 was 16.6 kcal $mol^{-1}$ higher than that of Cu–LCC-12, suggesting that Cu–LCC-4,4 is a less stable copper(ii) complex. High-resolution mass spectrometry (HRMS) confirmed the formation of monometallic copper biguanide complexes, with Cu–LCC-12 being the most stable (Fig. 2c and Extended Data Fig. 3c). LCC-12 did not form stable complexes with other divalent metal ions (Extended Data Fig. 3d). A reduction in the UV absorbance of LCC-12 upon addition of copper(ii) chloride indicated complex formation at low micromolar

concentrations. This was confirmed by the appearance of coloured solutions characteristic of metal complexes (Extended Data Fig. 3e,f). Notably, even at a 1,000-fold lower dose, LCC-12 antagonized the induction of CD86 and CD80 in aMDMs more potently than metformin (Fig. 2d). The effect of LCC-4,4 used at 10 μM was moderate, consistent with the reduced capacity of this analogue to form a complex with copper(ii). As reported for metformin[29], LCC-12 induced AMPK phosphorylation, albeit at a much lower concentration, suggesting that phenotypes induced by metformin are linked to copper(ii) targeting (Extended Data Fig. 3g).

Next, we evaluated the effect of LCC-12 on other cell types that can upregulate CD44 upon exposure to specific biochemical stimuli. LCC-12 interfered with the activation of dendritic cells and T lymphocytes and the expression of several cell surface molecules on alternatively activated macrophages (Extended Data Fig. 4a,b). By contrast, LCC-12 did not interfere with the activation of neutrophils, a process that is not marked by CD44 upregulation. Copper signalling has previously been linked to cancer progression[30,31,32]. Human non-small cell lung carcinoma cells and mouse pancreatic adenocarcinoma cells undergoing epithelial–mesenchymal transition (EMT)—a cell biology programme that can promote the acquisition of the persister cancer cell state and metastasis[12,33]—were characterized by CD44 upregulation and increased cellular copper. Consistently, LCC-12 interfered with EMT, as shown by the levels of the epithelial marker E-cadherin, mesenchymal markers vimentin and fibronectin, the EMT transcription factors Slug and Twist as well as the levels of pro-metastatic protein CD109 (Extended Data Fig. 4c,d). These data support a general mechanism involving copper that regulates cell plasticity.

Nanoscale secondary ion mass spectrometry (NanoSIMS) imaging of aMDMs revealed a subcellular localization of the isotopologue $^{15}N,^{13}C$-LCC-12 that overlapped with the signals of $^{197}Au$-labelled cytochrome *c*, suggesting that LCC-12 targets mitochondria (Extended Data Fig. 5a,b). Fluorescent in-cell labelling of the biologically active but-1-yne-containing analogue LCC-12,4 using click chemistry[34] gave rise to a cytoplasmic staining pattern that colocalized with cytochrome *c* (Fig. 2e,f). The mitochondrial staining of LCC-12,4 was reduced upon cotreatment with

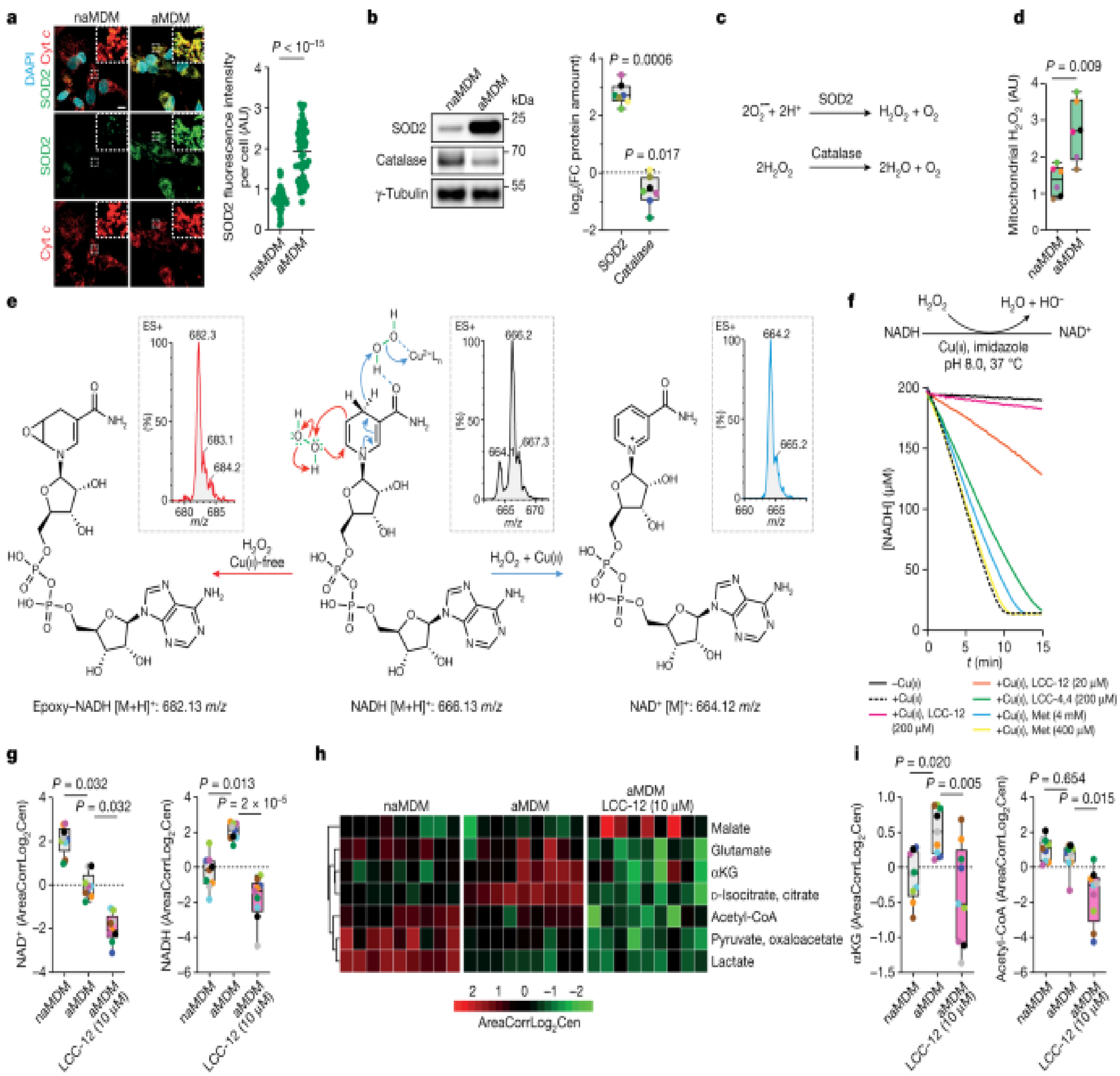
carbonyl cyanide chlorophenylhydrazone (CCCP), a small molecule that dissipates the inner mitochondrial proton gradient, indicating that LCC-12 accumulation in mitochondria is driven by its protonation state (Fig. 2g). Labelling alkyne-containing small molecules in cells requires a copper(i) catalyst generated in situ from added copper(ii) and ascorbate[34,35,36,37]. We investigated whether the mitochondrial copper(ii) content in aMDMs would allow in-cell labelling without the need to experimentally add a copper catalyst. Fluorescent labelling of LCC-12,4 used at a concentration of 100 nM, which is lower than the biologically active dose of LCC-12, occurred in aMDMs in the absence of added copper(ii) and a strong staining was observed only in aMDMs when ascorbate was used for labelling (Fig. 2h,i). Furthermore, the fluorescence intensity of labelled LCC-12,4 was reduced when a 100-fold molar excess of LCC-12 was used as a competitor (Extended Data Fig. 5c). These data support the existence of a druggable pool of chemically reactive copper(ii) in mitochondria. Consistent with this, levels of copper increased in mitochondria upon macrophage activation together with those of manganese (Fig. 2j and Extended Data Fig. 5d,e), whereas levels of copper in the endoplasmic reticulum and nucleus remained unaltered (Extended Data Fig. 5f,g). Notably, aMDMs were characterized by an increase of nuclear iron, hinting at an increased activity of αKG-dependent demethylases as previously shown in cancer cells undergoing EMT[19] (Extended Data Fig. 5f). LCC-12 treatment did not alter the total cellular and mitochondrial copper content of aMDMs, indicating that LCC-12 does not act as a cuprophore[38] (Extended Data Fig. 5h,i). By contrast, LCC-12 reduced the fluorescence of a mitochondrial copper(ii) probe[39] in aMDMs, supporting direct copper binding in mitochondria (Extended Data Fig. 5j). Notably, the mitochondrial metal transporters SLC25A3 and SLC25A37 were upregulated in aMDMs (Extended Data Fig. 5k). Knocking down the expression of these transporters or CD44 did not reduce labelled LCC-12,4 fluorescence (Extended Data Fig. 5l–n), whereas knocking down CD44 led to marked reduction of mitochondrial copper (Fig. 2k). This indicates that, unlike the proton gradient, mitochondrial copper does not drive mitochondrial accumulation of biguanides. As a control, labelling an alkyne-containing derivative of the copper(ii) chelator trientine, which did not exhibit a potent effect against macrophage activation, revealed nuclear accumulation, providing a rationale for the lack of biological activity of this

and potentially other copper-targeting drugs in this context (Extended Data Fig. 5o,p).

## Cu(ii) regulates NAD(H) redox cycling

Higher mitochondrial levels of manganese in aMDMs pointed to a functional role of the superoxide dismutase 2 (SOD2) in the context of macrophage activation. The amount of SOD2 protein increased in mitochondria upon activation, whereas the amount of catalase decreased (Fig. 3a,b). Mitochondrial hydrogen peroxide, a product of superoxide dismutase and substrate of catalase, increased accordingly (Fig. 3c,d). In cell-free systems, copper(ii) can catalyse the reduction of hydrogen peroxide by various organic substrates[40,41]. In the presence of copper(ii), NADH reacted with hydrogen peroxide to yield $NAD^+$, whereas the absence of copper(ii) yielded a complex mixture of oxidation products (Fig. 3e and Extended Data Fig. 6a). Consistently, copper(ii) favoured the conversion of 1-methyl-1,4-dihydronicotinamide (MDHNA), a structurally less complex surrogate of NADH, into 1-methylnicotinamide ($MNA^+$), whereas a product of epoxidation was formed preferentially in the absence of copper (Extended Data Fig. 6b,c). Thus, copper(ii) redirects the reactivity of hydrogen peroxide towards NADH. Under reaction conditions similar to those found in mitochondria, NADH was rapidly consumed to yield $NAD^+$ in the presence of copper(ii) (Fig. 3f and Extended Data Fig. 6d). This reaction was inhibited by LCC-12, whereas the effects of LCC-4,4 and metformin were marginal (Fig. 3f). Molecular modelling supported a reaction mechanism in which copper(ii) activates hydrogen peroxide, facilitating its reduction through the transfer of a hydride from NADH (Extended Data Fig. 6e). Copper(ii) acts as a catalyst that lowers the energy of the transition state with a geometry favouring this reaction. Molecular modelling also supported the inactivation of this reaction by biguanides through direct copper(ii) binding (Extended Data Fig. 6f).

**Fig. 3: Mitochondrial copper(ii) regulates NAD(H) redox cycling.**



**a**, Fluorescence microscopy of SOD2 in MDMs. Representative of $n = 4$ donors. At least 50 cells were quantified per donor. Scale bar, 10 µm. Two-sided unpaired *t*-test. Data are mean ± s.d. **b**, Representative western blots of SOD2 and catalase in MDMs ($n = 7$ donors). **c**, Regulation of $H_2O_2$ levels by SOD2 and catalase. **d**, Flow cytometry of mitochondrial $H_2O_2$ in MDMs ($n = 6$ donors). **e**, Reaction of NADH with $H_2O_2$ under copper(ii)-catalysed or copper-free conditions. Experimental mass spectrometry peaks and calculated masses of molecular ions are indicated. ES+, electrospray ionization mass spectrometry. **f**, Kinetics of NADH oxidation in the presence of $H_2O_2$ and copper(ii). Data are representative of $n = 3$ independent experiments. **g**, Metabolomics of $NAD^+$ and NADH of mitochondria from

MDMs treated with LCC-12 ($n = 9$ donors). AreaCorrLog$_2$Cen data correspond to raw areas, corrected for analytical bias using GRMeta R package, then corrected areas are log$_2$ transformed and centered on means. **h**, Metabolomics heat map highlighting metabolites whose biosynthesis is dependent on NAD(H) in MDMs treated with LCC-12 ($n = 9$ donors). **i**, Metabolomics of αKG and acetyl-CoA of MDMs treated with LCC-12 ($n = 9$ donors). **b**,**d**, Two-sided Mann–Whitney test. **g**,**i**, Kruskal–Wallis test with Dunn's post test. In graphs, each coloured dot represents an individual donor for a given panel.

Source data

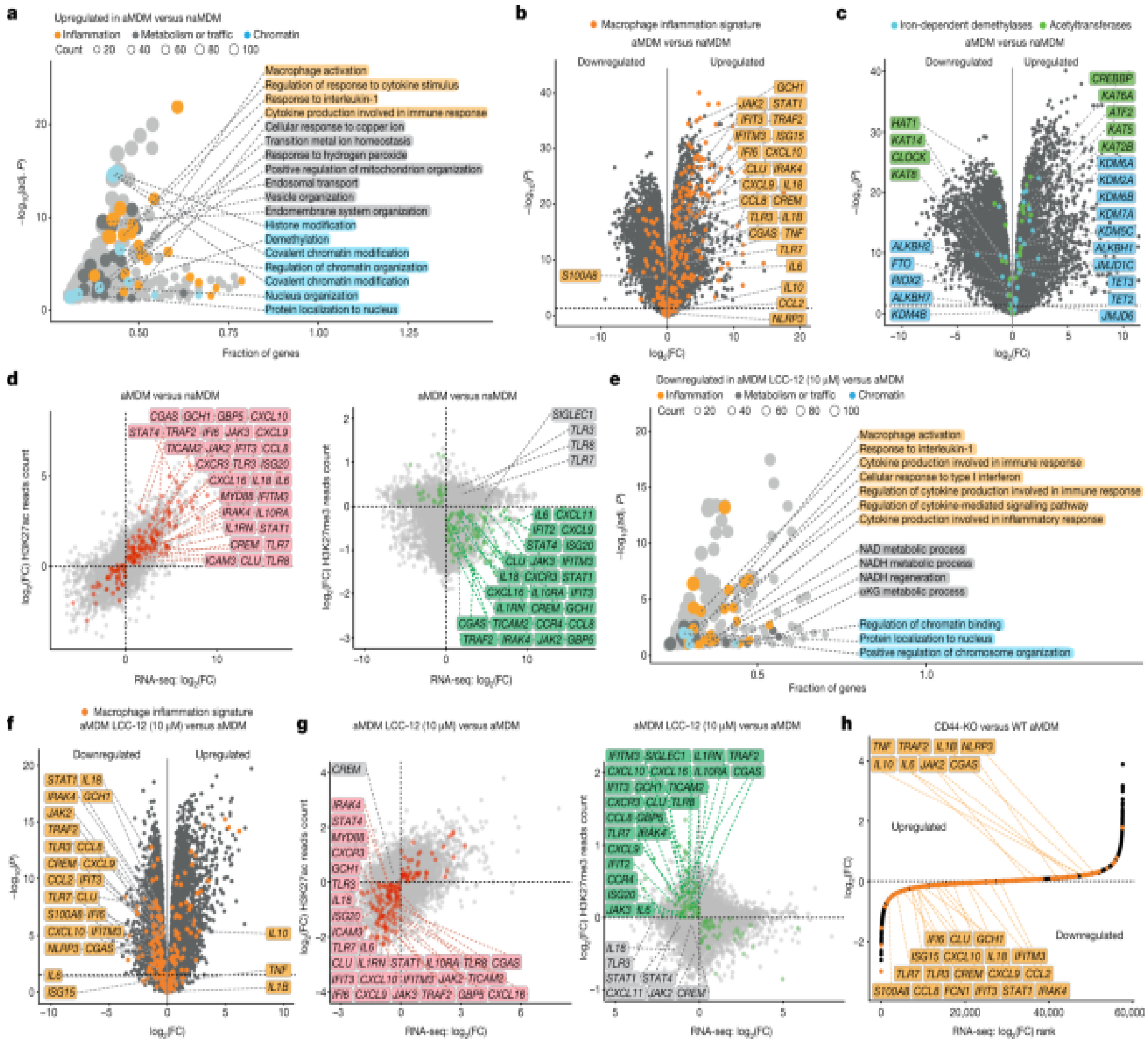
Mitochondrial NADH levels were higher and $NAD^+$ levels were lower in aMDMs compared with naMDMs, suggesting an enhanced activity of mitochondrial enzymes reliant on $NAD^+$ (Fig. 3g and Supplementary Table 1). Treating MDMs with LCC-12 during activation led to a reduction of NADH and $NAD^+$ (Fig. 3g and Supplementary Table 1). This suggests that copper(ii) catalyses the reduction of hydrogen peroxide by NADH to produce $NAD^+$ and that biguanides can interfere with this redox cycling, leading instead to other oxidation by-products (Fig. 3e). NADH and copper were found in mitochondria of aMDMs at an estimated substrate:catalyst ratio of 2:1, which is even more favourable for this reaction to take place than the 20:1 ratio used in the cell-free system (Extended Data Fig. 6g). Macrophage activation was accompanied by altered levels of several metabolites whose production depends on NAD(H) (Fig. 3h and Supplementary Table 2). LCC-12-induced metabolic reprogramming of aMDMs was marked by a reduction of αKG and acetyl-coenzyme A (acetyl-CoA) (Fig. 3i). LCC-12 also caused a reduction of extracellular lactate and accumulation of glyceraldehyde 3-phosphate in aMDMs consistent with the reduced activity of $NAD^+$-dependent glyceraldehyde 3-phosphate dehydrogenase (Extended Data Fig. 6h,i). Collectively, these data support the central role of mitochondrial copper(ii) in the maintenance of a pool of $NAD^+$ that regulates the metabolic state of inflammatory macrophages.

# Mitochondrial Cu(ii) regulates transcription

Transcription is co-regulated by chromatin-modifying enzymes, whose expression levels and recruitment at specific genomic loci shape gene expression. The turnover of specific enzymes such as iron-dependent demethylases and acetyltransferases relies on αKG and acetyl-CoA[42]. The finding that LCC-12 interfered with the production of these metabolites and opposed macrophage activation pointed to epigenetic alterations that affect the expression of inflammatory genes. We analysed the transcriptomes of aMDMs versus those of naMDMs by RNA sequencing (RNA-seq) (Supplementary Table 3) and compared them to transcriptomics data obtained from bronchoalveolar macrophages of individuals infected with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[43] and from human macrophages exposed in vitro to *Salmonella typhimurium*[44], *Leishmania major*[45] or *Aspergillus fumigatus*[46] (Supplementary Table 4). Gene ontology (GO) analysis revealed three groups of GO terms comprising upregulated genes, belonging to inflammation, metabolism and chromatin (Fig. 4a). Notably, the GO terms of these genes included endosomal transport, cellular response to copper ion, response to hydrogen peroxide and positive regulation of mitochondrion organization. Similar signatures were obtained for macrophages exposed to distinct pathogens (Extended Data Fig. 7a,b and Supplementary Table 5), as defined by GO terms and increased RNA amounts for genes involved in inflammation (Fig. 4b and Extended Data Fig. 7c). aMDMs exhibited upregulated genes encoding CD44, sorting nexin 9 (SNX9), a regulator of CD44 endocytosis, and metallothioneins (MT2A and MT1X) involved in copper transport and storage, whereas expression levels of *ATP7A* and *ATP7B* were downregulated (Supplementary Table 3). Genes involved in chromatin and histone modifications were upregulated in aMDMs and similar genes encoding iron-dependent demethylases and acetyltransferases were upregulated in bronchoalveolar macrophages from individuals infected with SARS-CoV-2, as well as in macrophages exposed to other pathogens (Fig. 4c and Extended Data Fig. 7d). These data indicate that distinct classes of pathogens trigger similar epigenetic alterations[47], leading to the inflammatory cell state. In aMDMs, variations in protein levels including increases in iron-dependent demethylases and acetyltransferases were consistent with the RNA-seq data (Extended Data Fig. 8a,b and Supplementary Table 6). Changes in levels of specific demethylases and acetyltransferases were associated with alterations of their targeted marks (Extended Data Fig. 8c,d). Chromatin

immunoprecipitation sequencing (ChIP–seq) revealed a global increase of the permissive acetyl marks H3K27ac, H3K14ac and H3K9ac together with a reduction of repressive methyl marks H3K27me3 and H3K9me2 at inflammatory gene loci, with consistent effects on the transcriptional profile of aMDMs (Fig. 4d, Extended Data Fig. 8e,f and Supplementary Table 7). LCC-12 treatment induced a downregulation of genes related to NAD(H) and αKG metabolism, regulation of chromatin and inflammation (Fig. 4e and Supplementary Table 8). Inactivating mitochondrial copper(ii) also promoted the downregulation of inflammatory genes at the RNA and protein levels (Fig. 4f, Extended Data Fig. 9a,b and Supplementary Table 3), reflecting a complex epigenetic reprogramming toward a distinct cell state (Extended Data Fig. 9c). LCC-12 treatment reduced H3K27ac, H3K14ac and H3K9ac and increased H3K27me3 and H3K9me2 levels (Extended Data Fig. 9d), which was associated with the downregulation of targeted inflammatory genes (Fig. 4g, Extended Data Fig. 9e,f and Supplementary Table 7). Thus, the LCC-12-induced decreases in αKG and acetyl-CoA were associated with a reduced activity of iron-dependent demethylases and acetyltransferases, respectively. Notably, knocking down expression of SOD2 or the mitochondrial copper transporter SLC25A3 reduced the inflammatory signature of macrophages (Extended Data Fig. 9g,h). Similarly, knocking out CD44 antagonized epigenetic programming of inflammation in aMDMs without adversely affecting the expression of other metal transporters (Fig. 4h, Extended Data Fig. 9i–k and Supplementary Table 9). Together, these data indicate that hydrogen peroxide is a driver of cell plasticity and that mitochondrial copper(ii) controls the availability of essential metabolic intermediates required for the activity of chromatin-modifying enzymes, which enables rapid transcriptional changes underlying the acquisition of distinct cell states.

**Fig. 4: Mitochondrial copper(ii) regulates the epigenetic states and transcriptional programmes of inflammatory macrophages.**




**a**, GO term analysis of upregulated genes in aMDMs ($n = 10$ donors). Adj. $P$, adjusted $P$ value. **b**, RNA-seq analysis of MDMs. Macrophage inflammatory signature genes are highlighted. The dashed line indicates an adjusted $P$ value of 0.05 ($n = 10$ donors). **c**, RNA-seq analysis of MDMs. Iron-dependent demethylase and acetyltransferase signature genes are highlighted. The dashed line indicates an adjusted $P$ value of 0.05 ($n = 10$ donors). **d**, Correlation for a representative donor of ChIP–seq reads count of histone marks in genes against RNA-seq of gene transcripts in MDMs ($n = 10$ donors). **e**, GO term analysis of genes in aMDMs ($n = 10$ donors) whose expression levels are downregulated upon treatment with LCC-12 ($n = 5$ donors). **f**, RNA-seq analysis of aMDMs ($n = 10$ donors) and MDMs treated with LCC-12 during activation ($n = 5$ donors). Macrophage inflammatory signature genes are highlighted. The dashed line indicates an

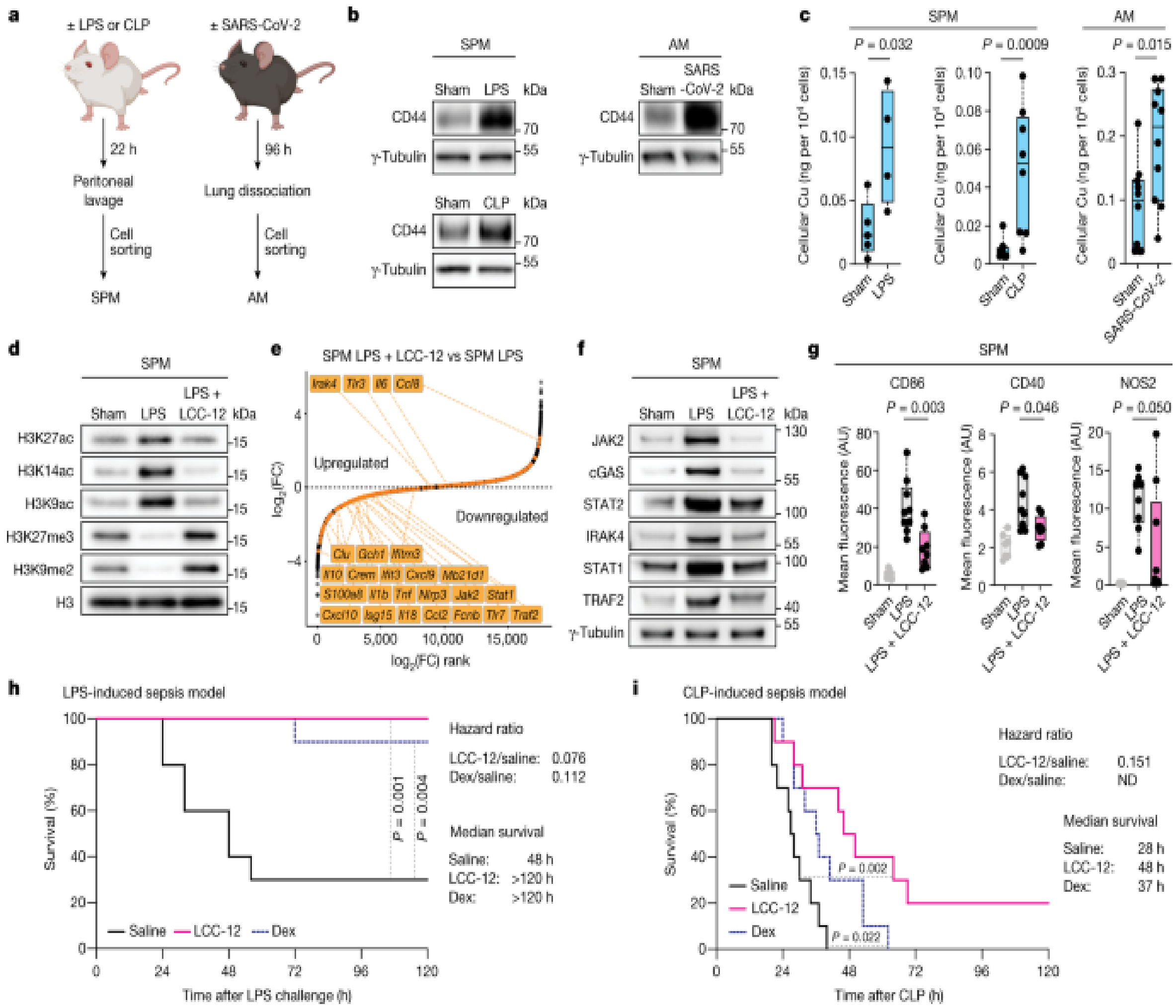
adjusted *P* value of 0.05. **g**, Correlation for a representative donor of ChIP–seq reads count of histone marks in genes against RNA-seq of gene transcripts in aMDMs (*n* = 10 donors) and MDMs treated with LCC-12 during activation (*n* = 5 donors). **h**, RNA-seq analysis of CD44-knockout (KO) and wild-type (WT) aMDMs. Representative of *n* = 4 donors. Gating strategy is shown in the Supplementary Information. Macrophage inflammatory signature genes are highlighted. The dashed line indicates an adjusted *P* value of 0.05. In **a**–**c**,**e**,**f**, Differential gene expression was assessed with the limma/voom framework. GO enrichment was assessed with the enrichGO method from clusterProfiler. *P* values were corrected for multiple testing with the Benjamini–Hochberg procedure.

# Cu(ii) inactivation reduces inflammation

We investigated the role of copper signalling in well-established mouse models of acute inflammation: (1) endotoxaemia induced by lipopolysaccharide (LPS), reflecting our mechanistic model of macrophage activation; (2) cecal ligation and puncture (CLP), which recapitulates the pathophysiology of subacute polymicrobial abdominal sepsis occurring in humans[48]; and (3) a model of viral infection, namely SARS-CoV-2. The inflammatory states of small peritoneal macrophages (SPMs) isolated from LPS and CLP mice models, and AMs isolated from SARS-CoV-2-infected mice, were characterized by upregulated CD44 and increased cellular copper (Fig. 5a–c). Effectors of the copper-signalling pathway, including HAS and SOD2 as well as specific epigenetic modifiers, were also upregulated in inflammatory macrophages (Extended Data Fig. 10a–c). Histone mark targets of these epigenetic modifiers were altered accordingly (Fig. 5d and Extended Data Fig. 10b,c). Intraperitoneal administration of LCC-12 in LPS-treated mice caused reductions in H3K27ac, H3K14ac and H3K9ac and increases in H3K27me3 and H3K9me2 (Fig. 5d), which were associated with reduced inflammation (Fig. 5e–g and Supplementary Table 10). In line with this result, intraperitoneal administration of LCC-12 fully protected mice from LPS-induced death and prevented the reduction of body temperature (Fig. 5h and Extended Data Fig. 10d), performing better than high-dose dexamethasone, which is used for the clinical management of acute inflammation. In CLP-induced sepsis, LCC-12 also increased the survival rate (Fig. 5i). LCC-12 administered by inhalation to SARS-CoV-2-

infected K18-hACE2 mice altered the expression of genes involved in the regulation of chromatin and downregulated the expression of inflammatory genes (Extended Data Fig. 10e,f and Supplementary Tables 11 and 12). Together, these data indicate that targeting mitochondrial copper(ii) interferes with the acquisition of the inflammatory state in vivo and confers therapeutic benefits.

**Fig. 5: Pharmacological inactivation of mitochondrial copper(ii) attenuates inflammation in vivo.**




**a**, Experimental setup to isolate SPMs and AMs. The gating strategy is shown in the Supplementary Information. **b**, Western blots of CD44 in inflammatory macrophages isolated from mice. Macrophages from 7–10 mice were pooled per condition. **c**, ICP-MS of cellular copper in SPMs or AMs from control mice (sham) and mice undergoing acute inflammation. LPS: sham ($n = 5$ mice), LPS-treated ($n = 4$ mice); CLP: sham ($n = 10$ mice),

with CLP ($n = 8$ mice); SARS-CoV-2: sham ($n = 10$ mice), SARS-CoV-2-infected ($n = 10$ mice). **d**, Western blots of histone marks in SPMs from mice treated with LPS and LCC-12. Macrophages from 4–7 mice pooled per condition. H3 is a sample processing control. **e**, Rank-order plot for RNA-seq of SPMs from mice treated with LPS and LCC-12. **f**, Western blots of proteins involved in inflammation in SPMs from mice treated with LPS and LCC-12. Macrophages from 4–7 mice were pooled per condition. **g**, Flow cytometry of SPMs from mice treated with LPS and LCC-12 ($n = 7$–9 mice). **h**, Kaplan–Meier survival curves of mice treated with LPS (20 mg $kg^{-1}$ per single dose; intraperitoneal injection; $n = 10$ mice) and LCC-12 (0.3 mg $kg^{-1}$ 2 h before challenge, then 24 h, 48 h, 72 h and 96 h after challenge; intraperitoneal injection; $n = 10$ mice) or dexamethasone (10 mg $kg^{-1}$ per dose 1 h before challenge; oral gavage; $n = 10$ mice). **i**, Kaplan–Meier survival curves of mice subjected to CLP and treated with LCC-12 (0.3 mg $kg^{-1}$ 4 h, 24 h, 48 h, 72 h and 96 h after CLP; intraperitoneal injection; $n = 10$ mice), dexamethasone (1.0 mg $kg^{-1}$ at time of CLP; intraperitoneal injection; $n = 10$ mice) or a saline solution (intraperitoneal injection; $n = 10$ mice). In **d**–**g**, LCC-12 (0.3 mg $kg^{-1}$) was injected 6 h after LPS and samples were collected 22 h after LPS. **c**,**g**, Two-sided Mann–Whitney test. **h**,**i**, Mantel–Cox log-rank test. Hazard ratio calculated using the Mantel–Haenszel method. ND, not determined.

Source data

# Discussion

CD44 has previously been linked to development, immune responses and cancer progression. Here we have shown that CD44 mediates the cellular uptake of specific metals, including copper, thereby regulating immune cell activation. We identified a chemically reactive pool of copper(ii) in mitochondria that characterizes the inflammatory state of macrophages. Our data support a cellular mechanism whereby the activation of hydrogen peroxide by copper(ii) enables oxidation of NADH to replenish the pool of $NAD^+$. Maintenance of this redox cycling is required for the production of key metabolites that are essential for epigenetic programming. In this context, copper(ii) acts directly as a metal catalyst, in contrast to its dynamic

metalloallosteric effect in other processes[30]. Transcriptomic shifts in macrophages exposed to distinct classes of pathogens substantiate the general nature of this mechanism.

We designed a dimer of biguanides that is able to inactivate mitochondrial copper(ii), thereby triggering metabolic and epigenetic reprogramming that reduces the inflammatory cell state and increases survival in preclinical models of acute inflammation. We have thus illustrated the pathophysiological relevance of this copper(ii)-triggered molecular chain of events. Acute inflammation is therefore reminiscent of a metabolic disease that can be rebalanced by targeting mitochondrial copper(ii) to restrict the generation of key metabolites required to initiate and maintain the inflammatory state (Extended Data Fig. 10g). LCC-12 selectively targets mitochondrial copper(ii), which is more abundant in the disease inflammatory cell state than in the basal state. This drug also interfered with the process of EMT in cancer cells, supporting a wider role for this copper-signalling pathway in the regulation of transcriptional changes beyond inflammation. Thus, CD44 may be characterized as a regulator of cell plasticity. Metformin exhibits positive effects on human health and is being studied as an anti-ageing drug[49,50]. However, investigation of its mechanism of action is hampered by its poor pharmacology resulting in low potency, which necessitates the administration of high doses. Thus LCC-12—which we rename ‘supformin’—exhibits improved biological and preclinical characteristics over metformin, making it a suitable drug-like small molecule for revealing novel mechanistic features of biguanides. Overall, our findings highlight the central role of mitochondrial copper(ii) as a regulator of cell plasticity and unveil a therapeutic strategy based on the control and fine-tuning of epigenetic cell states.

# Methods

## Ethics statement

Peripheral blood samples were collected from 128 healthy donors at Etablissement Français du Sang (EFS). The use of EFS blood samples from anonymous donors was approved by the Institut National de la Santé et de la Recherche Médicale committee. Written consent was obtained from all the

donors. Survival assessment using the LPS mouse model was conducted at Fidelta according to 2010/63/EU and national legislation regulating the use of laboratory animals in scientific research and for other purposes (Official Gazette 55/13). An institutional committee on animal research ethics (CARE-Zg) oversaw that animal-related procedures were not compromising the animal welfare. Flow cytometry, ICP-MS, western blotting and RNA-seq using the LPS mouse models were performed in accordance with French laws concerning animal experimentation (#2021072216346511) and approved by Institutional Animal Care and Use Committee of Université de Saint-Quentin-en-Yvelines (C2EA-47). All animal work using the CLP model was conducted in accordance with French laws concerning animal experimentation (#2021072216346511) and approved by Institutional Animal Care and Use Committee of Université de Saint-Quentin-en-Yvelines (C2EA-47). All animal work concerning RNA-seq on the SARS-CoV-2 model was performed within the biosafety level 3 facility of the Institut Pasteur de Lille, after validation of the protocols by the local committee for the evaluation of the biological risks and complied with current national and institutional regulations and ethical guidelines (Institut Pasteur de Lille/B59-350009). The experimental protocols using animals were approved by the institutional ethical committee Comité d'Ethique en Experimentation Animale (CEEA) 75, Nord-Pas-de-Calais. The animal study was authorized by the Education, Research and Innovation Ministry under registration number APAFIS#25517-2020052608325772v3. Animal work concerning cytometry, ICP-MS and western blotting on the SARS-CoV-2 model was performed within the biosafety level 3 facility of the University of Toulouse. This work was overseen by an Institutional Committee on Animal Research Ethics (license APAFIS#27729–2020101616517580 v3, Minister of Research, France (CEEA-001)), to ensure that animal-related procedures were not compromising the animal welfare.

## Antibodies

Antibodies are annotated below as follows. WB, western blot; FCy, flow cytometry; FM, fluorescence microscopy; NS, NanoSIMS; ChIP, ChIP–seq; Hu, used for human samples; Ms, used for mouse samples. Dilutions are indicated. Any antibody validation by manufacturers is indicated and can be

found on the manufacturers' websites. Our antibody validation by knockdown (KD) and/or KO strategies as described here for relevant antibodies is indicated. Primary antibodies: ALKBH1 (Abcam, ab195376, clone EPR19215, lot GR262105-2, WB 1:1,000, Hu, Ms, KO validated by manufacturer), AMP-activated protein kinase subunit alpha (AMPKα, Cell Signaling, 2532S, lot 21, WB 1:1,000, Hu), p-AMPKα phosphorylated on Thr172 (p-AMPKα, Cell Signaling, 2535S, lot 27, WB 1:1,000, Hu), ATF2 (Abcam, ab32160, clone E243, lot GR3430555-1, WB 1:1,000, Hu), ATF2 (Proteintech, 14834-1-AP, WB 1:1,000, Ms, KD/KO validated by manufacturer), ATP7A (Santa Cruz Biotechnology, sc-376467, clone D-9, lot J2821, WB 1:200, Hu), ATP7A (Novus Biologicals, NBP2-59376, clone S60-4, WB 1:1,000, Ms), ATP7B (Santa Cruz Biotechnology, sc-373964, clone A-11, lot I2719, WB 1:200, Hu, Ms), catalase (Cell Signaling, 12980T, clone D4P7B, lot 3, WB 1:1,000, Hu), CCL2 (also known as Mcp1) (Proteintech, 66272-1-Ig, clone 1B9F7, WB 1:1,000, Hu), CD11b-Pacific Blue (BioLegend, 101224, clone M1/70, lot B350151, B323654 and B323653, FCy 1:800, Ms), CD14-Krome Orange (Beckman Coulter, B01175, clone RMO/52, lot 200040, FCy 1:100, Hu), CD16-Pacific Blue (Beckman Coulter, B36292, clone 3G8, lot 200029, FCy 1:100, Hu), CD3 (BioLegend, 317326, clone OKT3, lot B372352, T cell activation, 2.5 μg $ml^{-1}$, Hu), CD25-BV711 (BioLegend, 302636, clone BC96, lot B281779, FCy 1:100, Hu), CD28 (BioLegend, 302934, clone CD28.2, lot 374639, T cell activation, 2.5 μg $ml^{-1}$, Hu), CD40-APC (BioLegend, 124612, clone 3/23, lot B309981, FCy 1:200, Ms), CD40-BV510 (BioLegend, 334329, clone 5C3, lot 312131, FCy 1:100, Hu), CD44 (Abcam, ab189524, lot GR320797-13, GR3314218-16, clone EPR18668, WB 1:30,000, Hu, Ms, KO/KD validated by us and manufacturer), CD44 (Thermo Fisher Scientific, 701406, clone 19H8L4, lot 1976318, FM 1:400, Hu), CD44-AF647 (Novus Biologicals, NB500-481AF647, clone MEM-263, lot 118753, FCy 1:100, Hu), CD44-AF647 (BioLegend, 103018, clone IM7, lot B317762, FCy 1:200, Ms), CD45-BV510 (BioLegend, 103138, clone 30-F11, lot B362964, B322199, B333193, FCy 1:200, Ms), CD64-FITC (BioLegend, 399505, clone S18012C, lot 308498, FCy 1:100, Hu), CD66b-PE/Cy7 (BioLegend, 305115, clone G10F5, lot 283925, FCy 1:100, Hu), CD69-PerCP (BioLegend, 310928, clone FN50, lot B290414, FCy 1:100, Hu), CD80-AF700 (BD Biosciences, 561133, clone 307.4, lot 1060235, FCy 1:100, Hu), CD83-PE (BioLegend, 305307, clone HB15e, lot B303073, FCy 1:100, Hu),

CD86-PE (BioLegend, 105007, clone GL-1, lot B318893, FCy 1:200, Ms), CD86-PE/Cy7 (BD Biosciences, 561128, clone 2331 (FUN-1), lot 1309531, FCy 1:33, Hu), CD109 (Santa Cruz Biotechnology, sc-271085, clone C-9, lot E1018, WB 1:200, Hu), CD109 (Biotechne, AF7717-SP, WB 1:1,000, Ms), CD163-PE (BD Biosciences, 556018, clone GHI/61, lot 9143793, FCy 1:100, Hu), CD170 (Siglec-F)-PEeFluor 610 (eBioscience, 61-1702-80, clone 1RNM44N, lot 2472220 and 2152352, FCy 1:200, Ms), cGAS (Cell Signaling, 15102, clone D1D3G, lot 4, WB 1:1,000, Hu), cGAS (Cell Signaling, 31659, clone D3O8O, lot 3, WB 1:1,000, Ms), CLOCK (Proteintech, 18094-1-AP, WB 1:1,000, Hu, KD/KO validated by manufacturer), CLOCK (Abcam, ab3517, WB 1:1,000, Ms), CTR1 (Abcam, ab129067, clone EPR7936, lot GR3414582-4 and GR81444-2, WB 1:1,000, Hu, KD validated by us), CTR2 (Novus Biologicals, NBP1-05199SS, WB 1:1,000, Hu), CTR2 (Novus Biologicals, NBP1-85512, lot R05901, FM 1:400, Hu), CTR2 (Biorybt, orb182668, lot BR2373, WB 1:1,000, Hu) COX IV (Abcam, ab16056, lot GR320655-1, FM 1:400, Hu), CREM (Proteintech, 12131-1-AP, WB 1:1,000, Hu), cytochrome *c* (Cell Signaling, 12963S, clone 6H2.B4, lot 1 and 2, FM 1:400, NS 1:400, Hu), DMT1 (Abcam, ab55735, clone 4C6, lot GR3243346-1, WB 1:1,000, Hu, KD validated by us), *Drosophila* spike-in antibody (Active Motif, 61686, lot 23521010, ChIP 50 ng per condition), E-cadherin (Cell Signaling, 3195, clone 24E10, lot 15, WB 1:1,000, Hu), E-cadherin (BD Biosciences, 610181, clone 36, lot 7187865, WB 1:1,000, Ms), F4/80-BV605 (BioLegend, 123133, clone BM8, lot B362524, B309659, B331465 and B339746, FCy 1:100, Ms), F4/80-PE (TONBO, TNB50-4801-U100, clone BM8.1, lot C4801060619503, FCy 1:100, Ms), Fibronectin (Sigma-Aldrich, F0791, clone IST-3, lot 026M4781V, WB 1:1,000, Hu, Ms), FTO (Proteintech, 27226-1-AP, WB 1:1,000, Hu, Ms, KD/KO validated by manufacturer), H3 (Cell Signaling, 9715S, lot 23, FM, WB 1:1,000, Hu, Ms), H3K4me3 (Diagenode, C15410003-50, lot A8034D, FM 1:400, Hu, dot blot validation by manufacturer), H3K9ac (Cell Signaling, 9649S, clone C5B11, lot 13, FM, WB, ChIP 6 µl per $10^6$ cells, Hu, Ms, validated with SimpleChIP Enzymatic Chromatin IP by manufacturer), H3K9me2 (Cell Signaling, 4658S, clone D84B4, lot 10, FM 1:400, WB 1:1,000, ChIP 6 µl per $10^6$ cells, Hu, Ms, validated with SimpleChIP Enzymatic Chromatin IP by manufacturer), H3K9me3 (Cell Signaling, 13969S, clone D4W1U, lot 3, FM 1:400, Hu, validated with SimpleChIP Enzymatic Chromatin IP by manufacturer),

H3K14ac (Cell Signaling, 7627S, clone D4B90, lot 6, FM, WB 1:1,000, ChIP 6 μl per $10^6$ cells, Hu, Ms, validated with SimpleChIP Enzymatic Chromatin IP by manufacturer), H3K27ac (Cell Signaling, 8173S, clone D5E4, lot 8, FM, WB 1:1,000, ChIP 6 μl per $10^6$ cells, Hu, Ms, validated with SimpleChIP Enzymatic Chromatin IP by manufacturer), H3K27me3 (Cell Signaling, 9733S, clone C36B11, lot 19, FM 1:400, WB 1:1,000, ChIP 6 μl per $10^6$ cells, Hu, Ms, validated with SimpleChIP Enzymatic Chromatin IP by manufacturer), H3K36me2 (Abcam, ab9049, lot GR3258133-1, FM 1:400, Hu), hyaluronan synthase 1 (HAS1, Novus Biologicals, NBP1-51635, clone 3E10, lot 141031, WB 1:1,000, Hu), HAS1 (Sigma-Aldrich, SAB4300848, lot 492637613, WB 1:1,000, Ms), hyaluronan synthase 2 (HAS2, Abcam, ab140671, clone 4E7, lot GR3212928-2, WB 1:1,000, Hu), HAS2 (Santa Cruz Biotechnology, sc-514737, clone A-7, WB 1:200, Ms), hyaluronan synthase 3 (HAS3, Abcam, ab154104, lot GR113715-12, WB 1:1,000, Hu), HAS3 (Proteintech, 15609-1-AP, WB 1:1,000, Ms, KO/KD validated by us and manufacturer), HAT1 (Proteintech, 11432-1-AP, WB 1:1,000, Hu, Ms, KO/KD validated by us and manufacturer), 5-hydroxymethylcytosine (5hmC, Active Motif, 39069, lot 23720003, FM 1:400, Hu, dot blot validated by manufacturer), I-A/I-E-AF700 (BioLegend, 107622, clone M5/114.15.2, lot B313251, FCy 1:400, Ms), IRAK4 (Cell Signaling, 4363T, lot 5, WB 1:1,000, Hu, Ms), JAK2 (Cell Signaling, 3230T, clone D2E12, WB 1:1,000, lot 13, Hu, Ms), JMJD6 (Abcam, ab64575, lot GR3441511-1, WB 1:1,000, Hu, Ms), KAT2B/PCAF (Cell Signaling, 3378T, clone C14G9, lot 2, WB 1:1,000, Hu, Ms, validated with SimpleChIP Enzymatic Chromatin IP by manufacturer), KAT3A (also known as CREBBP) (Abcam, ab2832, lot GR3360262-6, WB 1:1,000, Hu, Ms), KAT5 (also known as Tip60) (Santa Cruz Biotechnology, sc-166323, clone C-7, lot 166323, WB 1:200, Hu, Ms), KAT6A (also known as MOZ) (Santa Cruz Biotechnology, sc-293283, clone 4D8, lot F0420, WB 1:200, Hu), KAT6A (Invitrogen, PA5-103467, lot XH3653004, WB 1:1,000, Ms), KAT8 (also knon as MOF) (Proteintech, 13842-1-AP, KD/KO validated by manufacturer, WB 1:1,000, Hu, Ms), KDM2A (Abcam, ab191387, clone EPR18602, lot GR3330146-4, WB 1:1,000, Hu, Ms, KO validated by manufacturer), KDM5A (also known as Jarid1a) (Cell Signaling, 3876S, clone D28B10, lot 5, WB 1:1,000, Hu, Ms), KDM5B (also known as Jarid1b) (Cell Signaling, 3273T, lot 3, WB 1:1,000, Hu), KDM5B (Abcam, ab181089, clone EPR12794, WB 1:1,000, Ms, KO validated by

manufacturer), KDM5C (also known as Jarid1c) (Cell Signaling, 5361, clone D29B9, lot 1, 5361, WB 1:1,000, Hu, Ms), KDM6A (also known as UTX) (Cell Signaling, 33510S, clone D3Q1l, lot 4, WB 1:1,000, Hu, Ms), KDM6B (Abcam, ab169197, WB 1:1,000, Hu), KDM6B/JMJD3 (Cell Signaling, 3457, WB 1:1,000, Ms), KDM7A (Invitrogen, PA5-96987, lot UI2838718, WB 1:1,000, Hu, Ms), Ly6C–PerCP/Cy5.5 (BioLegend, 128012, clone HK1.4, lot B363119 and B310463, FCy 1:200, Ms), Ly6G–PE-Cy7 (BioLegend, 127618, clone 1A8, lot B288785 and B351626, FCy 1:200, Ms), Ly6G–AF647 (BioLegend, 127610, clone 1A8, lot B2559839, FCy 1:200, Ms), lysosome-associated membrane protein 2 (LAMP2) (Abcam, ab25631, clone H4B4, FM 1:400, Hu), NOS2–APC (eBioscience, 17-5920-82, clone CXNFT, lot 2154045, FCy 1:100, Ms), SLC25A3 (Santa Cruz Biotechnology, sc-376742, clone F-1, lot H2313, WB 1:200, FM 1:100, Hu, KO validated by us), SLC25A3 (Abcam, ab89117, lot 1015892-1, FM 1:400, Hu), SLC25A37 (MyBiosource, MBS9210193, clone ID: RB24153, lot SA100524AR, WB 1:1,000, FM 1:400, Hu, KD validated by us), Slug (Cell Signaling, 9585S, clone C19G7, lot 6, WB 1:1,000, Hu, Ms), STAT1 (Cell Signaling, 14994T, clone D1K9Y, lot 8, WB 1:1,000, Hu, Ms, validated with SimpleChIP Enzymatic Chromatin IP by manufacturer), STAT2 (Abcam, ab32367, clone Y141, lot GR3294792-5, WB 1:1,000, Hu, Ms, KO validated by manufacturer), SOD2 (Abcam, ab13534, lot GR3345921-12 and GR33618-52, FM 1:400, WB 1:1,000, Hu, Ms, KD validated by us), Tet methylcytosine dioxygenase 2 (TET2, Abcam, ab94580, lot GR3243631-2, WB 1:1,000, Hu, Ms), Tet methylcytosine dioxygenase 3 (TET3, Abcam, ab139311, lot GR3314447-1, WB 1:1,000, Hu, Ms), TRAF2 (Cell Signaling, 4724T, clone C192, lot 2, WB 1:1,000, Hu, Ms), TFR1 (Invitrogen, 13-6800, clone H68.4, lot VJ313549, WB 1:1,000, Hu, KD validated by us), TFR1–APC–AF750 (Beckman Coulter, A89313, clone YDJ1.2.2, lot 200060, FCy 1:100, Hu), γ-tubulin (Sigma-Aldrich, T5326, clone GTU-88, Source 0000128065, WB 1:1,000, Hu, Ms, enhanced validation by manufacturer), Twist (Santa Cruz Biotechnology, sc-81417, clone Twist2C1a, lot J2213, WB 1:200, Ms), vimentin (Cell Signaling, 5741S, clone D21H3, lot 8, WB 1:1,000, Hu, Ms, KO validated by previous users according to manufacturer's website). Secondary antibodies: Alexa Fluor 488 anti-rabbit (Invitrogen, A-11070, lot 2161039, FM 1:1,000, Hu), Alexa Fluor 594 anti-mouse (Invitrogen, A-11032, lot 1826426, FM 1:1,000, Hu), Alexa Fluor 594 anti-rabbit (Invitrogen,

A11072, lot 1985650, FM 1:1,000, Hu), Alexa Fluor 647 anti-mouse (Invitrogen, A21237, lot 1743738, FM 1:1,000, Hu), Alexa Fluor 647 anti-rabbit (Invitrogen, A21246, lot 2418503, FM 1:1,000, Hu), donkey anti-rabbit IgG-h+I HRP-conjugated (Bethyl Laboratories, A120-108P, lot 12 and 13), goat anti-mouse IgG h+I HRP-conjugated (Bethyl Laboratories, A90-116P, lot 41 and 44), 10 nM gold-nanoparticle-loaded anti-mouse (Abcam, ab27241, lot GR274015-2, NS 1:200, Hu).

## Primary cells

Peripheral blood samples were collected from 128 healthy donors at EFS. The use of EFS blood samples from anonymous donors was approved by the committee of INSERM (Institut National de la Santé et de la Recherche Médicale). Written consent was obtained from all the donors. Pan monocytes were isolated by negative magnetic sorting using microbeads according to the manufacturer's instructions (Miltenyi Biotec, 130-096-537) and cultured immediately in the presence of cytokines to trigger in vitro differentiation as described in 'Cell culture'. Cells were used fresh without prior freezing. Typically, cells were collected by incubation with 1× PBS with 10 mM EDTA at 37 °C and then scraped, unless stated otherwise. Primary non-small cell lung circulating cancer cells were obtained from Celprogen (36107-34CTC) and cultured as described in 'Cell culture'. Primary macrophages from in vivo mouse models (LPS, CLP and SARS-CoV-2) were isolated and processed as described in 'LPS-induced sepsis model', 'CLP-induced sepsis model' and 'SARS-CoV-2-induced acute inflammation model'.

## Cell culture

This study was carried out using primary monocytes obtained from peripheral blood samples of 128 distinct human donors from EFS. Pan monocytes were isolated by negative magnetic sorting using microbeads according to the manufacturer's instructions (Miltenyi Biotec, 130-096-537), and cultured in RPMI 1640 supplemented with GlutaMAX (Gibco, 61870010), 10% fetal bovine serum (FBS, Eurobio Scientific, CVFSVF00-01). To generate MDMs, pan monocytes were treated with granulocyte-macrophage colony-stimulating factor (GM-CSF, Miltenyi Biotec, 130-093-

866, 100 ng ml$^{-1}$) for 5 days. Subsequently, LPS (InvivoGen, tlrl-3pelps, 100 ng ml$^{-1}$) and interferon-γ (IFNγ, Miltenyi Biotec, 130-096-484, 20 ng ml$^{-1}$) were added to the medium for 24 h to generate aMDMs. CD4$^+$ T cells were isolated from peripheral blood samples by negative magnetic sorting using microbeads according to the manufacturer's instructions (Miltenyi Biotec, 130-096-533) and cultured in RPMI 1640 supplemented with glutamine and 10% fetal bovine serum. Subsequently, CD3/CD28 antibodies were added to the medium for 48 h to generate activated CD4$^+$ T cells. CD8$^+$ T cells were isolated from peripheral blood samples by negative magnetic sorting using microbeads according to the manufacturer's instructions (Miltenyi Biotec, 130-096-495) and cultured in RPMI 1640 supplemented with glutamine and 10% fetal bovine serum. Subsequently, CD3/CD28 antibodies were added to the medium for 48 h to generate activated CD4$^+$ T cells. To generate anti-inflammatory macrophages, pan monocytes were treated with macrophage colony-stimulating factor (M-CSF, Miltenyi Biotec, 130-096-492, 100 ng ml$^{-1}$) for 5 days and subsequently, IL-4 (Miltenyi Biotec, 130-093-921, 20 ng ml$^{-1}$) was added to the medium for 24 h. Neutrophils were isolated from peripheral blood samples. Red cells in whole blood samples were lysed (eBioscience RBC lysis buffer, 00-4300-54). The remaining cells were cultured in RPMI 1640 supplemented with glutamine, and 2% human serum. Subsequently, LPS (2 μg ml$^{-1}$) was added for 1 h to generate activated neutrophils. The neutrophil population was determined by flow cytometry using FSC, SSC and CD15 surface staining. To generate dendritic cells, pan monocytes were treated with GM-CSF (100 ng ml$^{-1}$) and IL-4 (10 ng ml$^{-1}$) for 5 days. Subsequently, LPS (100 ng ml$^{-1}$) was added to the medium to generate activated dendritic cells (aDC). FC1242 mouse pancreatic cancer cells were a generous gift from the Tuveson laboratory (Cold Spring Harbor Laboratory) and were cultured in Dulbecco's Modified Eagle Medium GlutaMAX (DMEM, Gibco, 61965059) supplemented with 10% FBS (Gibco, 10270-106) and penicillin–streptomycin mixture (BioWhittaker/Lonza, DE17-602E). Cells were treated with TGF-β (Miltenyi Biotec, 130-095-066, 10 ng ml$^{-1}$) for 6 days. Primary non-small cell lung circulating cancer cells (Celprogen, 36107-34CTC, lot 219411, sex: female) were grown using stem cell complete media (Celprogen, M36102-29PS) until the third passage. These cells were grown in stem cell

ECM T75-flasks (Celprogen, E36102-29-T75) and ECM 6-well plates (Celprogen, E36102-29-6Well) and treated with TGF-β for 3 days.

MDMs were activated with LPS and IFNγ and co-treated with ATTM (Sigma-Aldrich, 323446, 10 μM) or EDTA (Euromedex, EU0007, 500 μM), hyaluronate (Carbosynth, FH45321, 600-1000 kDa, 1 mg $ml^{-1}$), methylated hyaluronate (in-house, 1 mg $ml^{-1}$), LCC-12 (in-house, 10 μM), LCC-4,4 (in-house, 10 μM), metformin (1,1-dimethylbiguanide hydrochloride, Alfa Aesar, J63361, 10 mM), d-penicillamine (Sigma-Aldrich, P4875, 250 μM), trientine hydrochloride (Trien, Sigma-Aldrich, PHR1495-500MG, 200 μM) or anti-human CD44 therapeutic antibody (RG7356, Creative Biolabs, TAB-128CL, 10 μg $ml^{-1}$) for 24 h. aMDMs were treated with CCCP (Sigma-Aldrich, C2759, 10 μM), $^{15}N$,$^{13}C$-LCC-12 (in-house, 10 μM), LCC-12,4 (in-house, 100 nM) or trientine alkyne (in-house, 10 μM) for 3 h, by adding the reagents directly to the media. Dendritic cells were co-treated with LPS and LCC-12 (10 μM) for 24 h. $CD4^+$ and $CD8^+$ T cells were co-treated with CD3/CD28 and LCC-12 (50 μM) for 48 h. Anti-inflammatory macrophages were co-treated with IL-4 and LCC-12 (10 μM) for 24 h. Neutrophils were co-treated with LPS and LCC-12 (10 μM) for 1 h. Primary non-small cell lung circulating cancer cells and FC1242 cells were co-treated with TGF-β and LCC-12 (1 μM) for 3 days and 6 days, respectively.

## Flow cytometry

For human immune cells: cells were washed with ice-cold 1× PBS, incubated with Fc block (Human TruStain FcX, BioLegend, 422302, 1:20) for 15 min, subsequently incubated with antibodies for 20 min at 4 °C in 1× PBS/ 0.5% bovine serum albumin (BSA) and then washed before analysis using a flow cytometer (BD LSR Fortessa X-20). Cells were analysed with the corresponding antibody panels. Primary non-small cell lung circulating cancer cells and FC1245 cells were collected by trypsinization using Trypsin/EDTA (GIBCO, TRYPGIB01), washed with 1× PBS and antibody incubation was performed for 20 min at 4 °C in 1× PBS/10% FBS. Cells were washed and analysed using a flow cytometer (BD Accurie C6). For flow cytometry on cells from the LPS murine model, see the LPS-induced sepsis model paragraph. Flow cytometry analysis of mitochondrial $H_2O_2$ content: MDMs were activated with LPS and IFNγ, in the presence of

MitoPY1 (R&D Systems, 4428/10, 5 μM) for 24 h. Fluorescence was analysed by flow cytometry. Flow cytometry analysis with the mitochondrial Cu(ii) probe $M_{Cu}$-2[39]: naMDMs, aMDMs and MDMs that were activated with LPS and IFNγ and co-treated with LCC-12 (10 μM) were incubated with $M_{Cu}$-2 (5 μM) for 1h. Fluorescence was analysed by flow cytometry. The data were analysed with FlowJo software v. 10.8.2.

## ICP-MS

Glass vials equipped with Teflon septa were cleaned with nitric acid 65% (VWR, Suprapur, 1.00441.0250), washed with ultrapure water (Sigma-Aldrich, 1012620500) and dried. Cells were collected and washed twice with 1× PBS. Cells were then counted using an automated cell counter (Entek) and transferred in 200 μl 1× PBS or ultrapure water to the cleaned glass vials. The same volume of 1× PBS or ultrapure water was transferred into separate vials for the background subtraction, at least in duplicate per experiment. Mitochondria, nuclei and endoplasmic reticula were extracted as described in 'Isolation of mitochondria' from a pre-counted population of cells. Samples were lyophilized using a freeze dryer (CHRIST, 2–4 LDplus). Samples were subsequently mixed with nitric acid 65% and heated at 80 °C overnight in the same glass vials closed with a lid carrying a Teflon septum. Samples were then cooled to room temperature and diluted with ultrapure water to a final concentration of 0.475 N nitric acid and transferred to metal-free centrifuge vials (VWR, 89049-172) for subsequent mass spectrometry analyses. Amounts of metals were measured using an Agilent 7900 ICP-QMS in low-resolution mode, taking natural isotope distribution into account. Sample introduction was achieved with a micro-nebulizer (MicroMist, 0.2 ml $min^{-1}$) through a Scott spray chamber. Isotopes were measured using a collision-reaction interface with helium gas (5 ml $min^{-1}$) to remove polyatomic interferences. Scandium and indium internal standards were injected after inline mixing with the samples to control the absence of signal drift and matrix effects. A mix of certified standards was measured at concentrations spanning those of the samples to convert count measurements to concentrations in the solution. Values were normalized against cell number.

## Western blotting

MDMs were treated as indicated and then washed with 1× PBS. For MDMs, proteins were solubilized in 2× Laemmli buffer containing benzonase (VWR, 70664-3, 1:100). Extracts were incubated at 37 °C for 1 h and heated at 94 °C for 10 min, and quantified using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). SPMs and AMs from the LPS, CLP and SARS-CoV-2 mouse models were isolated by flow cytometry as described in the relevant flow cytometry section. Due to low cell number count, SPMs were collected in 1× PBS, which was subsequently freeze dried in Eppendorf tubes. Dried material was solubilized in 2× Laemmli buffer containing benzonase. AMs were pelleted and the cell pellets were solubilized in 2× Laemmli buffer containing benzonase. Extracts from SPMs and AMs were then incubated at 37 °C for 1 h and heated at 94 °C for 10 min. Proteins extracts from SPMs and AMs were quantified using a Qubit (Invitrogen) and a Qubit protein quantification assay (Invitrogen, Q33212). For the LPS model, SPMs were pooled for 8 sham mice, for 4 LPS-treated mice and for 6 LPS and LCC-12-treated mice. For the CLP model, SPMs were pooled from 8 sham mice and 7 mice subjected to CLP. For the SARS-CoV-2 model, AMs were pooled for 10 sham mice and for 10 SARS-CoV-2-infected mice. Protein lysates were resolved by SDS–PAGE (Invitrogen sure-lock system and NuPAGE 4–12% Bis-Tris precast gels). In a typical experiment from MDMs or cancer cells 10–20 μg of total protein extract was loaded per lane in 2× Laemmli buffer containing bromophenol blue. For in vivo isolated SPM and AMs where protein amounts were limited, typically 1 μg of total protein extract was loaded (more protein lysate was loaded in a new experiment if the antibody could not recognize a specific band at these protein amounts). On each gel a size marker was run (3 μl PageRuler or PageRuler plus, Thermo Scientific, 26616 or 26620 and 17 μl 2× Laemmli buffer) in parallel. Proteins were then transferred onto nitrocellulose membranes (Amersham Protran 0.45 μm) using a Trans-Blot SD semi-dry electrophoretic transfer cell (Bio-rad) using 1× NuPage transfer buffer (Invitrogen, NP00061) with 10% methanol. Membranes were blocked with 5% non-fat skimmed milk powder (Régilait) in 0.1% Tween-20/1× PBS for 20 min. Membranes were cut at the appropriate marker size to allow for the probing of several antibodies on the same membrane. Blots were then probed with the relevant primary antibodies in 5% BSA, 0.1% Tween-20/1× PBS or in 5% non-fat skimmed milk powder in 0.1% Tween-20/1× PBS at 4 °C overnight with gentle motion in a hand-sealed transparent plastic bag.

Membranes were washed with 0.1% Tween-20/1× PBS three times and incubated with horseradish peroxidase conjugated secondary antibodies (Jackson Laboratories) in 5% non-fat skimmed milk powder, 0.1% Tween-20/1× PBS for 1 h at room temperature and washed three times with 0.1% Tween-20/1× PBS. Antigens were detected using the SuperSignal West Pico PLUS (Thermo Scientific, 34580) and SuperSignal West Femto (Thermo Scientific, 34096) chemiluminescent detection kits. For blotting proteins on the same membranes, stripping buffer (0.1 M TRIS pH 6.8, 2% SDS w/v, 0.1 M β-mercaptoethanol) was used for 30 min and membranes were washed with 0.1% Tween-20/1× PBS and reblotted subsequently as described above. Signals were recorded using a Fusion Solo S Imaging System (Vilber). For histone marks, H3 was run as a sample processing control on a separate gel in parallel and is displayed in the respective panels. γ-tubulin served as loading control on the same gels and is not displayed in the respective panels. Band quantifications were performed with FIJI 2.0.0-rc-69/1.52n using pixel intensity normalized against the signal of γ-tubulin. All full scans of blots are displayed in the Supplementary Information.

## Fluorescence microscopy

Isolated monocytes were plated on coverslips, differentiated and activated as described in 'Cell culture'. For fluorescent detection of hyaluronate and Lys-Cu, live cells were treated with hyaluronate–FITC (800 kDa, Carbosynth, YH45321, 0.1 mg $ml^{-1}$) and Lys-Cu (in-house, 20 μM, 1 h)[26] for 1 h in the presence or absence of hyaluronidase (Sigma-Aldrich, H3884, 0.1 mg $ml^{-1}$). Hyaluronate–FITC and hyaluronidase were solubilized together in medium for 2 h at 37 °C before adding to the cells. Cells were then washed three times with 1× PBS, fixed with 2% paraformaldehyde in 1× PBS for 12 min and then washed 3 times with 1× PBS. For antibody staining, cells were then permeabilized with 0.1% Triton X-100 in 1× PBS for 5 min and washed 3 times with 1× PBS. Subsequently, cells were blocked in 2% BSA, 0.2% Tween-20/1× PBS (blocking buffer) for 20 min at room temperature. Cells were incubated with the relevant antibody in blocking buffer for 1 h at room temperature, washed 3 times with 1× PBS and were incubated with secondary antibodies for 1 h. Finally, coverslips were washed 3 times with 1× PBS and mounted using VECTASHIELD containing DAPI (Vector Laboratories, H-1200-10). Fluorescence images were acquired using a

Deltavision real-time microscope (Applied Precision). 40×/1.4NA, 60×/1.4NA and 100×/1.4NA objectives were used for acquisitions and all images were acquired as *z*-stacks. Images were deconvoluted with SoftWorx (Ratio conservative—15 iterations, Applied Precision) and processed with FIJI 2.0.0-rc-69/1.52n. Fluorescence intensity is displayed as arbitrary units (AU) and is not comparable between different panels. Colocalization quantification was calculated using FIJI 2.0.0-rc-69/1.52n. Histone quantification was performed using FIJI 2.0.0-rc-69/1.52n by delineating the nuclei using DAPI fluorescence, and calculating the mean fluorescence intensity normalized by area.

## Small molecule labelling using click chemistry

aMDMs on coverslips were treated with LCC-12,4 (in-house, 100 nM, 3 h) in absence or presence of CCCP or LCC-12 competitor (10 µM, 3 h), fixed and permeabilized as indicated in the fluorescence microscopy paragraph. Mitotracker (Invitrogen, M22426) was added to live cells for 45 min before fixation. For in-cell-labelling of trientine, live cells were incubated with trientine alkyne (in-house, 10 µM, 3 h). The click reaction cocktail was prepared using the Click-iT EdU Imaging kit (Invitrogen, C10337) according to the manufacturer's protocol. In a typical experiment we mixed 50 µl of 10× Click-iT reaction buffer with 20 µl of $CuSO_4$ solution, 1 µl Alexa Fluor–azide, 50 µl reaction buffer additive (sodium ascorbate) and 379 µl ultrapure water to reach a final volume of 500 µl. For variations as indicated in the figures, reactions were performed with or without $CuSO_4$ and ascorbate. Coverslips were incubated with the click reaction cocktail in the dark at room temperature for 30 min, then washed three times with 1× PBS. Immunofluorescence was then performed as described in 'Fluorescence microscopy'.

## NanoSIMS imaging

MDMs were grown on coverslips and activated to obtain aMDMs as described in 'Cell culture'. Cells were treated with 10 µM $^{15}N,^{13}C$-LCC-12 for 3 h. Subsequently, cells were washed twice with 1× PBS, once with 0.1 M cacodylate buffer (LFG Distribution, 11653) and then fixed with 2% paraformaldehyde in 0.1 M cacodylate buffer for 20 min. Then, cells were

washed three times with 0.1 M cacodylate buffer for 5 min and permeabilized with 0.1% Triton X-100 in 0.1 M cacodylate buffer for 5 min. Subsequently, cells were washed three times with 0.1 M cacodylate buffer and blocking buffer (2% BSA, 0.1% Tween in 0.1 M cacodylate buffer) was added for 20 min. Primary antibody (1:400) was added for 1 h in blocking buffer. Then, cells were washed three times with 0.1 M cacodylate buffer and the 10 nM gold-nanoparticle-loaded secondary antibody (1:50) was added in blocking buffer for 1 h. Cells were washed three times with 0.1 M cacodylate buffer and treated with 1% $OsO_4$ (Electron Microscopy Sciences, 19152) in 0.1 M cacodylate buffer for 1 h. Coverslips with samples were washed three times for 10 min with Milli-Q water. Subsequently, cells were dehydrated sequentially with ethanol solutions for 10 min each: 50%, 70%, 2× 90%, 3× 100% (dried over molecular sieves, Sigma-Aldrich, 69833). Samples were then coated with a 1:1 mixture of resin (Electron Microscopy Sciences, dodecenylsuccinic anhydride, 13710, methyl-5-norbornene-2,3-dicarboxylic anhydride, 19000, DMP-30, 13600 and LADD research industries: LX112 resin, 21310) and dry ethanol for 1 h. Then, samples were embedded in pure resin for 1 h. Embedding capsules (Electron Microscopy Sciences, 69910-10) were filled with resin, inverted onto the cover slides and placed in an oven at 56 °C for 24 h. Sections 0.2 µm in thickness were prepared using a Leica Ultracut UCT microtome. Sample sections were deposited onto a clean silicon chip (Institute for Electronic Fundamentals/CNRS and University Paris Sud) and dried upon exposure to air before being introduced into the NanoSIMS-50 ion microprobe (Cameca). A $Cs^+$ primary ion was employed to generate negative secondary ion from the sample surface. The probe steps over the image field and the signal of selected secondary ion species were recorded pixel-by-pixel to create 2D images. Image of $^{12}C^{14}N^-$ was recorded to provide the anatomic structure of the cells, while the one of $^{31}P^-$ highlights the location of cell nucleus. The cellular distribution of $^{15}N$ label was imaged by measuring the excess in $^{12}C^{15}N^-$ to $^{12}C^{14}N^-$ ratio with respect to the natural abundance level (0.0037), and the one for antibody with gold staining targeting mitochondria was performed by detecting directly $^{197}Au^-$ ions. When detecting $^{12}C^{15}N^-$ ion, appropriate mass resolution power was required to discriminate abundant $^{13}C^{14}N^-$ isobaric ions (with an $M/\Delta M$ of 4,272). For each image recording process, multiframe acquisition mode was applied and hundreds of image planes were recorded. The overall acquisition time

corresponding to the $^{15}N$ image was 12 h and 6 h 30 min for the $^{197}Au$ image. During image processing with FIJI (2.0.0-rc-69/1.52n), the successive image planes were properly aligned using TomoJ plugin[51], so as to correct the slight primary beam shift during long hours of acquisition. A summed image was then obtained with improved statistics. Further, for the $^{12}C^{15}N^{-}$ to $^{12}C^{14}N^{-}$ ratio map, an HSI (Hue-Saturation-Intensity) colour image was generated using OpenMIMS for display with increased significance[52]. The hue corresponds to the absolute $^{15}N/^{14}N$ ratio value, and the intensity at a given hue is an index of the statistical reliability.

## RNA interference

Human primary monocytes were transfected with Human Monocyte Nucleofector kit (Lonza, VPA-1007) according to the manufacturer's instructions. $5\times 10^6$ monocytes were resuspended in 100 µl nucleofector solution with 200 pmol of ON-TARGETplus SMARTpool siRNA or negative control siRNA (Qiagen, 1027310) before nucleofection with Nucleofector II (Lonza). Cells were then immediately removed and incubated overnight with 5 ml of prewarmed complete RPMI medium (Gibco). The following day, GM-CSF was added to the medium. The sequences of the SMARTpools used are detailed in the Supplementary Information.

## Genome editing

CRISPR knockout was performed using the following strategy. Human primary monocytes were transfected with Human Monocyte Nucleofector kit (Lonza, VPA-1007). Five million monocytes were resuspended in 100 µl nucleofector solution with a 100 pmol CAS9 (Dharmacon, CAS12206)/200 pmol CD44 (Dharmacon, SQ-009999-01-0010) single guide RNA (sgRNA) mix. The Cas9–CD44 mix was incubated for 10 min at 37 °C before nucleofection with Nucleofector II (Lonza). Cells were then immediately removed and incubated overnight with 5 ml of prewarmed complete RPMI medium (Gibco) and the following day, GM-CSF was added to the medium. At day 5, the cells were activated with LPS (100 ng $ml^{-1}$, 24 h) and IFNγ (20 ng $ml^{-1}$, 24 h). At day 6, cells were sorted for $CD44^{-}$ versus $CD44^{+}$ populations with BD FACSAria. The sorting strategy and the sequences

used for CD44 sgRNA (Edit-R Human Synthetic CD44, set of 3, target sequences) are detailed in the Supplementary Information.

## Bright-field microscopy and digital photographs

Bright-field images were acquired using a CKX41 microscope (Olympus) and cellSens Entry imaging software (Olympus). Digital images were taken with an iPhone 11 Pro (Apple).

## Isolation of mitochondria

Mitochondria were isolated using the Qproteome Mitochondria Isolation Kit (Qiagen, 37612) according to the manufacturer’s protocol. Cells were washed and centrifuged at 500*g* for 10 min and the supernatant was removed. Cells were then washed with a solution of 0.9% NaCl (Sigma-Aldrich, S7653-250G) and resuspended in ice-cold lysis buffer and incubated at 4 °C for 10 min. The lysate was then centrifuged at 1,000*g* for 10 min at 4 °C and the supernatant carefully removed. Subsequently, the cell pellet was resuspended in disruption buffer. Complete cell disruption was obtained by using a dounce homogenizer (mitochondria for ICP-MS) or a blunt-ended needle and a syringe (mitochondria for metabolomics). The lysate was then centrifuged at 1,000g for 10 min at 4 °C and the supernatant transferred to a clean tube. The supernatant was then centrifuged at 6,000*g* for 10 min at 4 °C to obtain mitochondrial pellets.

## Isolation of nuclei

Nuclei were isolated using the Nuclei EZ Prep (Sigma-Aldrich, NUC101-1KT) according to the manufacturer’s instructions. In brief, cells were treated as indicated and collected upon scraping and counted. Subsequently, cells were washed twice with 1× PBS and lysed with 1 ml ice-cold Nuclei EZ lysis buffer for 5 min on ice. The suspension was centrifuged at 500*g* for 5 min at 4 °C. Resulting nuclei were washed with Nuclei EZ lysis buffer and centrifuged to generate a pellet of isolated cell nuclei.

## Isolation of endoplasmic reticula

Endoplasmic reticula were isolated using the Endoplasmic Reticulum Enrichment Extraction Kit (Novus Biologicals, NBP2-29482) according to the manufacturer's instructions. In brief, 500 µl of 1× isosmotic homogenization buffer followed by 5 µl of 100× PIC were added to a pellet of $10^6$ cells. The resulting suspension was centrifuged at 1,000*g* for 10 min at 4 °C. The supernatant was transferred to a clean centrifuge tube and centrifuged at 12,000*g* for 15 min at 4 °C. The floating lipid layer was discarded. The supernatant was centrifuged in a clean centrifuge tube using an ultracentrifuge at 90,000*g* for 1 h. The resulting pellet contained the total endoplasmic reticulum fraction (rough and smooth).

## NMR spectroscopy

$^1$H NMR, $^{13}$C NMR and $^{15}$N NMR spectra were recorded on a 500 MHz Bruker spectrometer at 298 K or 310 K, and chemical shifts δ are expressed in ppm using the residual solvent signals as internal standard. To measure the interaction of hyaluronate tetrasaccharide with copper, portions of 0.25 mol equivalent of a solution of $CuCl_2$ in $D_2O$ (8.6 mg in 599 µl $D_2O$) were added to a 2 mM solution of hyaluronate tetrasaccharide (TCI Chemicals, H1284) in $D_2O$ (1.0 mg hyaluronate tetrasaccharide in 600 µl $D_2O$) up to 1 mol equivalent into an NMR tube. Then, a drop of trifluoroacetic acid (TFA, Alfa Aesar, A12198) was added. In a separate NMR tube, a drop of TFA was added to a copper-free 2 mM solution of hyaluronate tetrasaccharide in $D_2O$. To measure oxidation of NADH into $NAD^+$, $^1$H NMR were recorded and tubes were prepared as follows: NADH (Sigma-Aldrich, N4505-100MG, 200 µM) or $NAD^+$ (Sigma-Aldrich, N0632-1G, 200 µM), imidazole (10 mM), $CuSO_4$ (10 µM) were added as indicated to an NMR tube containing sodium phosphate buffer in $D_2O$ (10 mM, pD 8.4, 981 µl) or $D_2O$. $^1$H NMR were recorded at $t_0$. $H_2O_2$ (19 µl of 100× diluted solution of 32.3% (w/w%) in $H_2O$) was added to the NMR tubes and $^1$H NMR were recorded after 1 h.

## Chemical synthesis

Products were purified on a preparative HPLC Quaternary Gradient 2545 equipped with a Photodiode Array detector (Waters) fitted with a reverse phase column (XBridge BEH C18 OBD Prep column 5 µm 30×150 mm). NMR Spectra were run in DMSO-$d_6$, methylene chloride-$d_2$ or methanol-$d_4$ at 298 K unless stated otherwise. $^1$H NMR spectra were recorded on Bruker spectrometers at 400 or 500 MHz. Chemical shifts δ are expressed in ppm using residual non-deuterated solvent signals as internal standard. The following abbreviations are used: ex, exchangeable; s, singlet; d, doublet; t, triplet; td, triplet of doublets; m, multiplet. The $^{13}$C NMR spectra were recorded at 100.6 or 125.8 MHz, and chemical shifts δ are expressed in ppm using deuterated solvent signal as internal standard. The purity of final compounds, determined to be >98% by UPLC-MS, and low-resolution mass spectra (LRMS) were recorded on a Waters Acquity H-class equipped with a Photodiode array detector and SQ Detector 2 (UPLC-MS) fitted with a reverse phase column (Acquity UPLC BEH C18 1.7 µm, 2.1x50 mm). HRMS were recorded on a Thermo Scientific Q-Exactive Plus equipped with a Robotic TriVersa NanoMate Advion.

**LCC-12**

Dicyandiamide (Alfa Aesar, A10451, 500 mg, 5.94 mmol), 1,12-diaminododecane (Alfa Aesar, A04258, 500 mg, 2.50 mmol) and $CuCl_2$ (Sigma-Aldrich, 22.201-1, 249 mg, 1.85 mmol) were suspended in water (6 ml) in a sealed tube and stirred for 1 h at room temperature, then heated at 80 °C for 48 h. The resulting pink mixture was filtered and the solid was resuspended in water (10 ml). $H_2S$, generated from dropwise addition of 37% aqueous (aq.) HCl (Supelco, 1.00317.100) on FeS (~100 mesh powder, Alfa Aesar, 17422), was passed into the mixture until it turned black. The black mixture was filtered, and the filtrate was acidified to pH 5 with a 1 M aq. solution of HCl. The solvent was evaporated under reduced pressure. LCC-12 was purified by preparative HPLC ($H_2O$:acetonitrile:formic acid, 95:5:0.1 to 0:100:0.1) to afford LCC-12 di-formic acid salt as a white powder (280 mg, 24%). $^1$H NMR (500 MHz, DMSO-$d_6$) δ: 8.80–8.08 (m, 2H, ex), 8.47 (s, 2H, formate), 7.60–6.78 (m, 12H, ex), 3.14–2.93 (m, 4H), 1.55–1.34 (m, 4H), 1.32–1.18 (m, 16H) ppm. $^{13}$C NMR (125.8 MHz, DMSO-$d_6$) δ: 167.2 (formate), 160.3, 159.2, 41.3, 29.5 (3C), 29.2, 26.8 ppm.

HRMS (ESI+) *m/z*: calculated for $C_{16}H_{38}N_{10}$ $[M+2H]^{2+}$ 185.1635, found 185.1637.

**LCC-4,4**

Bis-(cyanoguanidino)butane (200 mg, 0.90 mmol) and butylamine hydrochloride (TCI Chemical, B0710, 197 mg, 1.80 mmol) were mixed together in a sealed tube and heated at 150 °C without solvent for 4 h. After cooling to room temperature, the mixture was taken up in ethanol and a large excess of ethyl acetate was added. The white precipitate was filtered and purified by preparative HPLC ($H_2O$:acetonitrile:formic acid, 100:0:0.1 to 50:50:0.1) to give the LCC-4,4 di-formic acid salt as a white powder (130 mg, 31%). $^{1}H$ NMR (400 MHz, DMSO-$d_6$) δ: 8.82–7.77 (m, 4H, ex), 8.46 (s, 2H, formate), 7.36–6.80 (m, 8H, ex), 3.16–2.99 (m, 8H), 1.58–1.37 (m, 8H), 1.37–1.20 (m, 4H), 0.86 (t, $J$ = 6.8 Hz, 6H) ppm. $^{13}C$ NMR (100.6 MHz, DMSO-$d_6$) δ: 167.3 (formate), 159.1 (2C), 40.9 (2C), 31.6, 26.8, 20.0, 14.1 ppm. HRMS (ESI+) *m/z*: calculated for $C_{16}H_{38}N_{10}$ $[M+2H]^{2+}$ 185.1635, found 185.1636.

The synthesis of bis-(cyanoguanidino)butane was adapted from a previously published procedure[53]. 1,4-diaminobutane (Acros Organics, 112120250, 554 mg, 5,68 mmol) was dissolved in water and stirred with 37% aq. HCl for 10 min at room temperature. The solvent was evaporated under reduced pressure and the resulting salt was suspended in butanol (5 ml) with sodium dicyanamide (Sigma-Aldrich, 178322, 1.09 g, 11.4 mmol) and stirred at 140 °C overnight. After filtration, the solid was washed with butanol and cold water and recrystallized from water to give the bis-(cyanoguanidino)butane as a white powder (350 mg, 27%). $^{1}H$ NMR (500 MHz, DMSO-$d_6$) δ: 7.62–6.12 (m, 6H, ex), 3.23–2.85 (m, 4H), 1.71–1.15 (m, 4H) ppm. $^{13}C$ NMR (125.8 MHz, DMSO-$d_6$) δ: 161.6, 118.8, 40.9, 26.8 ppm.

**LCC-12,4**

Bis-(cyanoguanidino)dodecane (227 mg, 0.60 mmol) and but-3-yne-1-amine hydrochloride (Enamine, EN300-76524, 126 mg, 1.20 mmol) were mixed

together in a sealed tube and heated at 150 °C without solvent for 4 h. After cooling to room temperature, the mixture was taken up in ethanol and a large excess of ethyl acetate was added slowly. The white precipitate was filtered and purified by preparative HPLC ($H_2O$:acetonitrile:formic acid, 95:5:0.1 to 40:60:0.1) to give LCC-12,4 di-formic acid salt as a white powder (102 mg, 30%). $^1$H NMR (500 MHz, DMSO-$d_6$) δ: 9.05–7.95 (m, 4H, ex), 8.47 (s, 2H, formate), 7.60–6.60 (m, 8H, ex), 3.29–3.16 (m, 4H), 3.11–3.00 (m, 4H), 2.84 (s, 2H), 2.39–2.29 (m, 4H), 1.49–1.38 (m, 4H), 1.33–1.18 (m, 16H) ppm. $^{13}$C NMR (125.8 MHz, DMSO-$d_6$) δ: 167.6 (formate), 159.7, 158.5, 82.6, 72.7, 41.3, 40.3, 29.5 (3C), 29.2, 26.8, 19.4 ppm. HRMS (ESI+) *m*/*z*: calculated for $C_{24}H_{46}N_{10}$ $[M+2H]^{2+}$ 237.1948, found 237.1947.

The synthesis of bis-(cyanoguanidino)dodecane was adapted from a previously published procedure[53]. 1,12-diaminododecane (500 mg, 2.5 mmol) was dissolved in a mixture of water and methanol and stirred with 37% aq. HCl for 10 min at room temperature. The solvent was evaporated under reduced pressure and the resulting salt was suspended in butanol (2.5 ml) with sodium dicyanamide (444 mg, 5.0 mmol) and stirred at 140 °C overnight. After filtration the solid was washed with butanol and cold water and recrystallized from a mixture of water:ethoxyethanol (2:1) to give the bis-(cyanoguanidino)dodecane as a white powder (365 mg, 44%). $^1$H NMR (400 MHz, DMSO-$d_6$) δ: 7.21–6.18 (m, 6H, ex), 3.11–2.94 (m, 4H), 1.49–1.34 (m, 4H), 1.33–1.14 (m, 16H) ppm. $^{13}$C NMR (100.6 MHz, DMSO-$d_6$) δ: 161.6, 118.8, 41.0, 29.4 (3C), 29.2, 26.7 ppm.

## Isotopically labelled LCC-12

$^{15}$N- and $^{13}$C-labelled dicyandiamide (Eurisotop, CNLM-9324-PK, 50 mg, 0.55 mmol), 1,12-diaminododecane (46.8 mg, 0.23 mmol) and $CuCl_2$ (31.4 mg, 0.23 mmol) were suspended in a sealed tube in water (0.6 ml) and stirred for 1 h at room temperature, then heated at 80 °C for 48 h. The resulting pink mixture was acidified with an aq. solution of HCl (2 M, 1 ml) until complete dissolution of the precipitate. The mixture was concentrated under reduced pressure and isotopically labelled LCC-12 was purified by preparative HPLC ($H_2O$:acetonitrile:formic acid, 95:5:0.1 to 73:27:0.1) to

give the isotopically labelled LCC-12 di-formic acid salt as a white powder (35 mg, 32%). $^{1}H$ NMR (500 MHz, DMSO-$d_6$) δ: 8.60–7.76 (m, 2H, ex), 8.48 (s, 2H, formate), 7.70–6.30 (m, 12H, ex), 3.12–2.95 (m, 4H), 1.53–1.35 (m, 4H), 1.33–1.18 (m, 16H) ppm. $^{13}C$ NMR (125.8 MHz, DMSO-$d_6$) δ: 167.0 (formate), 160.2, 159.2, 41.3, 29.5 (3C), 29.2, 26.8 ppm. $^{15}N$ NMR (50.7 MHz, DMSO-$d_6$) δ: 156.2, 84.1 (2N), 81.2 ppm. HRMS (ESI+) *m*/*z*: calculated for $C_{12}{}^{13}C_4H_{37}N_2{}^{15}N_8$ $[M+H]^+$ 381.3095, found 381.3093.

### Trientine alkyne

Under argon atmosphere, trientine dihydrochloride (Santa Cruz Biotechnology, sc-216009, 0.050 g, 0.228 mmol) was dissolved in anhydrous methanol at 0 °C followed by the addition of 4-(prop-2-ynyloxy)-benzaldehyde[54] (0.073 mg, 0.456 mmol) and molecular sieves 4 Å. The mixture was stirred for 3 h at room temperature, prior to the addition of $NaBH_3CN$ (0.026 g, 0.684 mmol) and stirred overnight at room temperature. Next, the reaction mixture was filtered, the filtrate was evaporated under reduced pressure and purified by preparative HPLC ($H_2O$:acetonitrile:formic acid, 95:5:0.1 to 73:27:0.1) to afford trientine alkyne as a white powder (20.7 mg, 21%). $^{1}H$ NMR (500 MHz, Methanol-$d_4$) δ: 7.40–7.33 (m, 4H), 7.06–6.99 (m, 4H), 4.74 (d, $J$ = 2.4 Hz, 4H), 3.97 (s, 4H), 3.00–2.91 (m, 10H), 2.87 (s, 4H). $^{13}C$ NMR (125.8 MHz, Methanol-$d_4$) δ: 159.4, 131.7 (2C), 128.7, 116.4 (2C), 79.6, 76.9, 56.7, 52.4, 47.7, 46.8 (2C). (ESI+) *m*/*z*: calculated for $C_{26}H_{35}N_4O_2$ $[M+H]^+$ 435.2755, found 435.2758.

### Lysosomal copper(ii) probe

The lysosomal copper(ii) probe Lys-Cu was synthesized as previously reported and spectral data are in agreement with the literature[26]. $^{1}H$ NMR (400 MHz, Methylene chloride-$d_2$) δ: 8.69–8.59 (m, 2H), 8.31 (dd, $J$ = 8.5, 1.2 Hz, 1H), 8.04 (d, $J$ = 1.5 Hz, 1H), 7.82–7.69 (m, 2H), 7.64 (dd, $J$ = 7.8, 1.7 Hz, 1H), 7.24 (d, $J$ = 7.7 Hz, 1H), 6.60 (d, $J$ = 8.8 Hz, 2H), 6.45 (d, $J$ = 2.5 Hz, 2H), 6.38 (dd, $J$ = 8.9, 2.6 Hz, 2H), 4.33 (t, $J$ = 6.8 Hz, 2H), 3.74 (s,

2H), 3.65–3.62 (m, 4H), 3.40–3.34 (m, 9H), 2.69 (t, $J$ = 6.9 Hz, 2H), 2.56 (s, 3H), 1.19 (t, $J$ = 7.1 Hz, 12H).

### Mitochondrial copper(ii) probe

The mitochondrial copper(ii) probe $M_{Cu}$-2 was synthesized as previously reported and spectral data are in agreement with the literature[39]. $^1$H NMR (400 MHz, DMSO-$d_6$) δ: 9.89 (s, 1H), 8.07–7.67 (m, 16H), 7.58–7.45 (m, 2H), 7.02–6.93 (m, 1H), 6.80 (d, $J$ = 2.4 Hz, 1H), 6.62 (d, $J$ = 2.3 Hz, 1H), 6.58–6.36 (m, 4H), 4.44 (s, 2H), 4.08 (t, $J$ = 6.0 Hz, 2H), 3.75–3.58 (m, 2H), 2.02–1.83 (m, 2H), 1.83–1.60 (m, 2H).

### Methylated hyaluronate

Methylated hyaluronate (meth-HA) was synthesized using modified published methods[25,55]. In brief, 1 g of sodium hyaluronate (600–1,000 kDa) was dissolved overnight in 200 ml of distilled water at room temperature. Then amberlite cation exchange resin ($H^+$) (10 g) was added to the solution and stirred for one day at room temperature. The resin was subsequently filtered off the solution. The resulting solution was then neutralized using tetrabutylammonium hydroxide (TBAOH) to obtain (tetrabutylammonium) TBA–hyaluronate as follows: TBAOH, diluted fivefold with water was added dropwise to the previously prepared hyaluronic acid solution until the pH reached 8. The solution was then freeze dried and the resulting TBA–hyaluronate was kept in a freezer until further use. 0.530 g of TBA–hyaluronate were dissolved in 10 ml of DMSO at 30 °C, then 1.5 ml of methyl iodide were added and the solution was kept at 30 °C overnight. The resulting mixture was slowly poured into 200 ml of ethyl acetate under constant agitation. The white precipitate obtained was filtered and washed four times with 100 ml of ethyl acetate and finally vacuum dried for 24 h at room temperature. Methylation of hyaluronate was confirmed by $^1$H NMR spectroscopy and size-exclusion chromatography. $^1$H NMR spectra were acquired on a Bruker 400 MHz spectrometer in $D_2O$ containing 0.125 M sodium deuteroxide to increase hyaluronate proton mobility, thus improving spectra resolution. $^1$H NMR spectrum displays

peaks of methyl NHCOMe at 1.79–1.86 ppm, as well as peaks of methyl groups OMe at 2.61, 2.78 ppm and COOMe at 3.22 ppm.

## Computational structural characterization of metformin-based Cu complexes

The starting structure for $Cu(Met)_2$ was based on the published X-ray structure[56]. The starting geometries for the other copper(ii) complexes were obtained via molecular dynamics conformation search (Gabedit[57], amber99 (ref. [58]) potential). For each complex, the ten geometries with the lowest energies resulting from the molecular dynamics search were reoptimized using MOPAC2016 (PM7 (ref. [59]), COSMO[60] water model). The geometries with the lowest energy for each complex were optimized at TPSSh/D3BJ/Def2-TZVP level using Orca 4.2.1 (ref. [61]). We performed a benchmark study based on the structure of $Cu(Met)_2$ using B3LYP[62], M062X[63], TPSSh, BHLYP[64,65] functionals with the Def2-TZVP[66] basis set using D3BJ[67] dispersion correction and the CPCM water solvation model. We have also used the BHLYP functional with the SVP basis set and the SMD water solvation model, which was recommended in the literature[68] for copper(ii) complexes.

## Energy calculation of copper(ii)-catalysed hydride transfer to $H_2O_2$ from NADH

The UBHLYP functional[64,65] was used associated with the SVP basis set[69,70] and the SMD solvation model[71,72] to represent an adequate method to describe the $[Cu(H_2O)_6]^{2+}$ species[68]. Thus, all structures (minima and transition states) were optimized using the Gaussian 16 set of programs at the UBHLYP/SVP level for all atoms (doublet spin state). The SMD solvation model (water) was applied during the optimization process. Thermal correction to the Gibbs free energy was computed at 310.15 K. Single points at the UMP2/SVP level were performed. The results presented are $\Delta G_{298}$ in kcal $mol^{-1}$. MDHNA was chosen as a NADH model to study the copper(ii)-catalysed hydride transfer to $H_2O_2$.

## HRMS of biguanide–metal complexes

HRMS solutions were prepared and injected without further dilution. Stock solution of LCC-12:$K_2CO_3$ (1:2) (100 µM) or metformin:$K_2CO_3$ (1:1) (200 µM) were prepared in analytical grade methanol. Stock solutions of metals, $CuCl_2.2H_2O$ (Acros Organics, 405840050, 100 µM), $MnCl_2$ (Sigma-Aldrich, 244589, 100 µM), $CaCl_2.2H_2O$ (Acros Organics, 423520250, 100 µM), $FeCl_2.xH_2O$ (Alfa Aesar, 12357, 100 µM), $MgCl_2.6H_2O$ (Prolabo, 25 108.295, 100 µM), $NiCl_2$ (Alfa Aesar, 53131, 100 µM) and $ZnCl_2$ (Sigma-Aldrich, 429430, 100 µM) were prepared in milliQ water. The HRMS solutions were prepared in methanol (80 µl) with LCC-12:$K_2CO_3$ (10 µl) and the relevant metal (10 µl) with a 1:1 ratio, or with metformin:$K_2CO_3$ (10 µl) and $CuCl_2.2H_2O$ (10 µl) with a 2:1 ratio.

## UV titration experiments

To a solution of LCC-12 (5 µM) in HEPES (10 mM), portions of 0.1 mol equivalent of a solution of $CuCl_2$ in HEPES (10 mM) were added up to 3 mol equivalent. UV spectra were recorded on an Analytik Jena UV/ VIS spectrophotometer specord 205 system at room temperature in the 200–1000 nm range using a micro cuvette (quartz Excellence Q 10 mm). All spectra were blanked against HEPES buffer.

## Copper-catalysed oxidation of NADH

The oxidation kinetics of NADH (Sigma-Aldrich, N4505) were followed by measuring the absorbance at 340 nm using a Cary 300 UV-Vis spectrometer. The measurements were recorded at 37 °C controlled with a Pelletier Cary temperature controller (Agilent Technologies). Stock solutions of NADH (1 mM), imidazole (Sigma-Aldrich, 56750, 100 mM), $CuSO_4$ (Sigma-Aldrich, 451657, 500 µM), LCC-12 (10 mM or 1 mM), metformin·HCl (Alfa Aesar, J63361, 100 mM or 10 mM) and LCC-4,4 (10 mM) were prepared in a 10 mM sodium phosphate buffer adjusted to pH 8.0. The concentration of $H_2O_2$ (Sigma-Aldrich, 16911, 32.3% wt in $H_2O$) was determined by titration with $KMnO_4$ and diluted 100 times in the phosphate buffer. In disposable

cuvettes, 1 ml of experimental solutions were prepared using sodium phosphate buffer and respective stock solutions to attain the final concentrations described Fig. 3f, by adding: NADH (200 µl), imidazole (100 µl), $CuSO_4$ (20 µl), LCC-12 (20 µl), LCC-4,4 (20 µl) or metformin (40 µl) as indicated, and $H_2O_2$ solution (19 µl) at $t_0$. The concentration of NADH was calculated from the measured absorbance at 340 nm and its molar extinction coefficient.

## Measurement of NADH concentrations

NADH absolute concentrations were measured using a fluorometric assay (Abcam, ab176723) according to the manufacturer's protocol. At least 500,000 cells were collected per condition. Floating cells were collected and adherent cells were washed with 1× PBS. Adherent cells were incubated with 1× PBS with 10 mM EDTA and then scraped and pooled together with the collected floating cells. Cells were subsequently washed with ice-cold 1× PBS and counted, then centrifuged at 1,500 rpm for 5 min and the supernatant discarded. The pellet was then resuspended in 100 µl lysis buffer (kit component) and incubated at 37 °C for 15 min. $NAD^+$ and NADH extraction solutions as well as $NAD^+$/NADH control solutions (kit components) were added and incubated at 37 °C for 15 min at a volume of 15 µl sample to 15 µl of the respective buffers (kit components). The reactions were stopped using 15 µl of respective buffers (kit components). Finally, 75 µl of $NAD^+$/NADH reaction mixture ($NAD^+$/NADH recycling enzyme mixture and sensor buffer, kit components) were added and the resulting mixtures incubated for 1 h at room temperature. Fluorescence intensities (excitation 540 nm; emission 590 nm) were recorded using a Perkin Elmer Wallac 1420 Victor2 Microplate Reader. Values were derived from the standard curve of each experiment and compared to the data obtained by mass spectrometry-based metabolomics, to calculate total and mitochondrial NADH concentrations.

## Quantitative metabolomics

In a typical experiment, 1.5 million cells were used for total extracts and 15 million cells for mitochondrial extracts. Cells were collected and the supernatant removed to generate the corresponding cell pellets.

Subsequently, pellets were dried and supplemented with 300 µl methanol, vortexed 5 min and centrifuged (10 min at 15,000*g*, 4 °C). Then, the upper phase of the supernatant was split into two parts: 150 µl were used for gas chromatography–mass spectrometry (GC–MS) experiment in microtubes and the remaining 150 µl were used for ultra high pressure liquid chromatography–mass spectrometry (UHPLC–MS). For the GC–MS aliquots, supernatants were completely evaporated from the sample. 50 µl of methoxyamine (20 mg $ml^{-1}$ in pyridine) were added to the dried extracts, then stored at room temperature in the dark for 16 h. The following day, 80 µl of *N*-methyl-*N*-(trimethylsilyl) trifluoroacetamide was added and final derivatization occurred at 40 °C for 30 min. Samples were then transferred into vials and directly injected for GC–MS analysis. For the UHPLC–MS aliquots, 150 µl were dried in microtubes at 40 °C in a pneumatically-assisted concentrator (Techne DB3). The dried UHPLC–MS extracts were solubilized with 200 µl of MilliQ water. Aliquots for analysis were transferred into liquid chromatography vials and injected into UHPLC–MS or kept at –80 °C until injection. Widely-targeted analysis of intracellular metabolites gas chromatography coupled to a triple-quadrupole mass spectrometer (QQQGC–MS): the GC–MS/MS method was performed on a 7890A gas chromatography (Agilent Technologies) coupled to a triple-quadrupole 7000C (Agilent Technologies) equipped with a High sensitivity electronic impact source operating in positive mode[73]. Peak detection and integration of the analytes were performed using the Agilent Mass Hunter quantitative software (B.07.01). Targeted analysis of nucleotides and cofactors by ion pairing ultra high performance liquid chromatography (UHPLC) coupled to a Triple Quadrupole (QQQ) mass spectrometer: targeted analysis was performed on a RRLC 1290 system (Agilent Technologies) coupled to a Triple Quadrupole 6470 (Agilent Technologies) equipped with an electrospray source operating in both negative and positive modes. Gas temperature was set to 350 °C with a gas flow of 12 l $min^{-1}$. Capillary voltage was set to 5 kV in positive mode and 4.5 kV in negative mode. Ten microlitres of sample were injected on a Column Zorbax Eclipse XDB-C18 (100 mm × 2.1 mm particle size 1.8 µm) from Agilent technologies, protected by a guard column XDB-C18 (5 mm × 2.1 mm particle size 1.8 µm) and heated at 40 °C by a pelletier oven. The gradient mobile phase consisted of water with 2 mM of dibutylamine acetate concentrate (DBAA) (A) and acetonitrile (B). Flow rate was set to 0.4 ml

min$^{-1}$ and an initial gradient of 90% phase A and 10% phase B, which was maintained for 3 min. Molecules were then eluted using a gradient from 10% to 95% phase B over 1 min. The column was washed using 95% mobile phase B for 2 min and equilibrated using 10% mobile phase B for 1 min and the autosampler was kept at 4 °C. Scan mode used was the MRM for biological samples. Peak detection and integration of the analytes were performed using Agilent Mass Hunter quantitative software (B.10.1). Pseudo-targeted analysis of intracellular metabolites by UHPLC coupled to a Q-Exactive mass spectrometer. Reversed phase acetonitrile method: the profiling experiment was performed with a Dionex Ultimate 3000 UHPLC system (Thermo Fisher Scientific) coupled to a Q-Exactive (Thermo Fisher Scientific) equipped with an electrospray source operating in both positive and negative modes and full scan mode from 100 to 1,200 *m*/*z*. The Q-Exactive parameters were: sheath gas flow rate 55 au, auxiliary gas flow rate 15 au, spray voltage 3.3 kV, capillary temperature 300 °C, S-Lens RF level 55 V. The mass spectrometer was calibrated with sodium acetate solution dedicated to low mass calibration. 10 µl of sample were injected on a SB-Aq column (100 mm × 2.1 mm particle size 1.8 µm) from Agilent Technologies, protected by a guard column XDB-C18 (5 mm × 2.1 mm particle size 1.8 µm) and heated at 40 °C by a pelletier oven. The gradient mobile phase consisted of water with 0.2% acetic acid (A) and acetonitrile (B). The flow rate was set to 0.3 ml min$^{-1}$. The initial condition was 98% phase A and 2% phase B. Molecules were then eluted using a gradient from 2% to 95% phase B for 22 min. The column was washed using 95% mobile phase B for 2 min and equilibrated using 2% mobile phase B for 4 min. The autosampler was kept at 4 °C. Peak detection and integration were performed using the Thermo Xcalibur quantitative software (2.1.)[73].

## Quantitative proteomics

Cells were grown and treated as indicated. Whole-cell extracts were collected by scraping after incubation with 1× PBS with 10 mM EDTA at 37 °C. After centrifugation at 1,500*g* for 5 min at 4 °C, cells were washed twice with ice-cold 1× PBS and lysed using lysis buffer (8 M urea, 200 mM $NH_4HCO_3$, cOmplete) for 1 h at 4 °C on a rotary wheel. After centrifugation at 20,000*g*, 4 °C for 20 min, supernatants that contain proteins were used for the global proteome analysis. In brief, the global proteome was

quantitatively analysed with a Orbitrap Eclipse mass spectrometer using a label-free approach. About 10 µg of total protein cell lysate were reduced by incubation with 5 mM dithiothreitol (DTT) at 57 °C for 30 min and then alkylated with 10 mM iodoacetamide for 30 min at room temperature in the dark. The samples were then diluted with 100 mM ammonium bicarbonate to reach a final concentration of 1 M urea and digested overnight at 37 °C with Trypsin:Lys-C (Promega, V5071) at a ratio of 1:50. Samples were then loaded onto a homemade C18 StageTips for desalting. Peptides were eluted from beads by incubation with 40:60 acetonitrile:water with 0.1% formic acid. Peptides were dried in a Speedvac and reconstituted in 10 µl 0.3% TFA prior to liquid chromatography-tandem mass spectrometry (LC–MS/MS) analysis. Samples of 4 µl were chromatographically separated using an RSLCnano system (Ultimate 3000, Thermo Fisher Scientific) coupled online to an Orbitrap Eclipse mass spectrometer (Thermo Fisher Scientific). Peptides were first loaded onto a C18-trapped column (75 µm inner diameter × 2 cm; nanoViper Acclaim PepMap 100, Thermo Fisher Scientific), with buffer A (2:98 MeCN:$H_2O$ with 0.1% formic acid) at a flow rate of 3 µl $min^{-1}$ over 4 min and then switched for separation to a C18 column (75 µm inner diameter × 50 cm; nanoViper C18, 2 µm, 100 Å, Acclaim PepMap RSLC, Thermo Fisher Scientific) regulated to a temperature of 50 °C with a linear gradient of 2 to 30% buffer B (100% MeCN and 0.1% formic acid) at a flow rate of 300 nl $min^{-1}$ over 211 min. MS1 data were collected in the Orbitrap (120,000 resolution; maximum injection time 60 ms; AGC $4 \times 10^5$). Charges states between 2 and 5 were required for MS2 analysis, and a 45 s dynamic exclusion window was used. MS2 scans were performed in the ion trap in rapid mode with HCD fragmentation (isolation window 1.2 Da; NCE 30%; maximum injection time 60 ms; AGC $10^4$). The identity of proteins was established from the UniProt human canonical database (UP000005640_9606) using Sequest HT through proteome discoverer (version 2.4) (Thermo Scientific). Enzyme specificity was set to trypsin and a maximum of two missed cleavage sites were allowed. Oxidized methionine, Methionine-loss, Methionine-loss-acetyl and N-terminal acetylation were set as variable modifications. Carbamidomethylation of cysteins were set as fixed modification. Maximum allowed mass deviation was set to 10 ppm for monoisotopic precursor ions and 0.6 Da for MS/MS peaks. The resulting files were further processed using myProMS v3.9.3 (https://github.com/bioinfo-pf-curie/myproms)[74]. For the false discovery rate

(FDR) calculation we used Percolator[75] and, this was set to 1% at the peptide level for the whole study. The label-free quantification was performed by peptide Extracted Ion Chromatograms (XICs) computed with MassChroQ version 2.2.21 (ref. [76]). For protein quantification, XICs from proteotypic peptides shared between compared conditions (TopN matching) with up to two missed cleavages and carbamidomethyl modifications were used. Median and scale normalization was applied on the total signal to correct the XICs for each biological replicate. To estimate the significance of the change in protein abundance, a linear model (adjusted on peptides and biological replicates) was performed and *P* values were adjusted with a Benjamini–Hochberg FDR procedure with a control threshold set to 0.05 and the proteins should have at least 3 peptides[75].

## Lactate quantification

Extracellular lactate was quantified using a fluorometric lactate assay (Abcam, ab65330) according to the manufacturer's instructions. The culture media of cells was collected and centrifuged at 500*g* for 5 min. The supernatant was subsequently centrifuged at 20,000*g* for 10 min. The supernatant was then deproteinized using 10 kD spin columns (Abcam, ab93349). A lactate standard was prepared by adding 5 µl of the 100 nmol $\mu l^{-1}$ lactate standard to 495 µl of lactate assay buffer. Subsequently, 1 ml of 0.01 nmol $\mu l^{-1}$ lactate standard was produced by diluting 10 µl of 1 nmol $\mu l^{-1}$ standard to 990 µl of lactate assay buffer. In a 96-well plate, standard samples of 0–0.1 nmol per well were added. A reaction mix of 46 µl lactate assay buffer, 2 µl probe and 2 µl enzyme mix was prepared for each well. For background measurements, 48 µl lactate assay buffer were mixed with 2 µl probe to obtain a background reaction mix. Fifty microlitres of each sample was added to a 96-well plate and either the reaction mix or the background reaction mix was added. Samples were incubated for 30 min at room temperature. Fluorescence intensities (excitation 540 nm; emission 590 nm) were recorded using a Perkin Elmer Wallac 1420 Victor2 Microplate Reader. Values were derived from the standard curve.

## Glyceraldehyde 3-phosphate quantification

Glyceraldehyde 3-phosphate (GA3P) was quantified using a fluorometric glyceraldehyde 3-phosphate assay kit (Abcam, ab273344) adapting the manufacturer's instructions. Cells were washed twice with 1× PBS and then collected into a centrifugation tube in 100 µl GA3P Assay Buffer. Samples were kept on ice for 10 min and then centrifuged at 10,000× g for 10 min. The supernatant was collected and then deproteinized using 10 kD spin columns (Abcam, ab93349). For each test sample, 10 µl of sample were added into three parallel wells in a white, flat bottom 96-well plate. A sample background control, an un-spiked and spiked sample were added to these three wells. The spiked sample contained 200 pmol of GA3P standard. 50 µl of GA3P assay buffer was added per well. For the assay blank, 50 µl GA3P assay buffer was added per well. To each background well, 50 µl of a reaction mix consisting of 46 µl GA3P assay buffer, 2 µl GA3P enzyme mix and 2 µl GA3P probe was added. To each remaining well 50 µl of a reaction mix was added, which consisted of 44 µl GA3P assay buffer, 2 µl GA3P developer, 2 µl GA3P enzyme mix and 2 µl GA3P probe. Fluorescence intensities (excitation 540 nm; emission 590 nm) were recorded using a Perkin Elmer Wallac 1420 Victor2 Microplate Reader. Several readings were performed at 1 min intervals. The final GA3P concentrations were calculated by subtracting background sample values.

## Luminex immunoassay

Cytokine levels were measured in cell culture supernatants using the V-Plex pro-inflammatory panel (MSD, K15049D-1). The kit was run according to the manufacturer's protocol and the chemiluminescence signal was measured on a Sector Imager 2400 (MSD).

## RNA-seq

RNAs were extracted from MDMs using the RNeasy mini kit (Qiagen, 74104). RNA sequencing libraries were prepared from 1 µg total RNA using the Illumina TruSeq Stranded mRNA library preparation kit (Illumina, 20020594), which allows strand-specific sequencing. A first step of polyA selection using magnetic beads was performed to allow sequencing of polyadenylated transcripts. After fragmentation, cDNA synthesis was performed and resulting fragments were used for dA-tailing followed by

ligation of TruSeq indexed adapters (Illumina, 20020492). Subsequently, polymerase chain reaction amplification was performed to generate the final barcoded cDNA libraries. Sequencing was carried out on a NovaSeq 6000 instrument from Illumina based on a 2× 100 cycles mode (paired-end reads, 100 bases). For RNA-seq on cells from the in vivo murine models, see the respective paragraphs. Raw sequencing reads were first checked for quality with Fastqc (0.11.8) and trimmed for adapter sequences with the trimGalore (0.6.2) software. Trimmed reads were then aligned on the human hg38 reference genome using the STAR mapper (2.6.1b), up to the generation of a raw count table per gene (GENCODE annotation v29). The bioinformatics pipelines used for these tasks are available online (rawqc v2.1.0: https://github.com/bioinfo-pf-curie/raw-qc, RNA-seq v3.1.4: https://github.com/bioinfo-pf-curie/RNA-seq). The downstream analysis was then restricted to protein-coding genes. Data from the literature[77] were converted into bulk by keeping cells annotated as macrophages and then summing the counts for each sample. Counts data from the literature[44] were downloaded from GEO under accession number GSE73502. Raw data from the literature[45,46] were downloaded from the NCBI Short Read Archive under records PRJNA528433 and PRJNA290995 and processed as described above. For *L. major*, we used data at 4 h post-infection; for *A. fumigatus*, we used data at 2 h post-infection. Counts were normalized using TMM normalization from edgeR (v 3.30.3)[78]. Differential expression was assessed with the limma/voom framework (v 3.44.3)[79]. The intra-donor correlation was controlled by using the duplicateCorrelation from limma. Genes with an adjusted $P$ value < 0.05 were labelled significant. Enrichment analysis from differentially expressed genes has been performed using the enrichGO function from clusterProfiler package v3.16.1.

## ChIP–seq

Cells were grown and treated as described. Cells were centrifuged at 1,500*g* for 5 min at room temperature. The pelleted cells were resuspended in medium, counted and crosslinked with 1% formaldehyde for 10 min at room temperature. Then, 2.5 M glycine was added to a final concentration of 0.125 M and incubated for 5 min at room temperature followed by centrifugation at 1,500*g* for 5 min at 4 °C. The pelleted cells were washed twice with ice-cold 1× PBS and collected by centrifugation at 1,500*g* for 5

min at 4 °C. Pellets were resuspended in lysis buffer A (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS, cOmplete) and incubated for 30 min on a rotating wheel at 4 °C. Next, lysates were centrifuged at 1,500*g* for 15 min at 8 °C to prevent SDS precipitation, and the supernatants were discarded. Pellets were then sheared in buffer B (25 mM Tris-HCl pH 8, 3 mM EDTA, 0.1% SDS, 1% Triton X-100, 150 mM NaCl, cOmplete) to approximately 200–600 bp average size using a Bioruptor Pico (Diagenode). After centrifugation at 20,000*g* at 4 °C for 15 min, supernatants containing sheared chromatin were used for immunoprecipitation. Twenty-five microlitres (10%) of sheared chromatin was used as input DNA to normalize sequencing data. As an additional control for normalization, spike-in chromatin from *Drosophila* (Active Motif, 53083) and a spike-in antibody were used. Chromatin immunoprecipitation (1 million cells per ChIP condition) was carried out using sheared chromatin and antibodies against specific histone marks, which were subsequently complexed to either Dynabeads Protein G-coated magnetic beads (Invitrogen, 10003D) for H3K27ac, H3K14ac, H3K9ac, H3K27me3 and H3K9me2. In brief, each antibody was mixed with 1 μg of spike-in antibody. Then, 22 μl magnetic beads were washed three times in ice-cold buffer C (20 mM Tris-HCl pH 8, 2 mM EDTA, 0.1% SDS, 1% Triton X-100, 150 mM NaCl, cOmplete) and incubated with the mixture of antibodies for 4 h, at room temperature on a rotating wheel in buffer C (494 μl). After spinning and removal of supernatants, beads were resuspended in 50 μl buffer C. This suspension was subsequently incubated with 250 μl sheared chromatin previously mixed with 50 ng of spike-in chromatin from *Drosophila* (250 μl chromatin of interest: 2.5 μl spike-in chromatin) at 4 °C on a rotating wheel overnight (16 h). After a spinning, supernatants were discarded and beads were successively washed in buffer C (twice), buffer D (20 mM Tris-HCl pH 8, 2 mM EDTA, 0.1% SDS, 1% Triton X-100, 500 mM NaCl), buffer E (10 mM Tris-HCl pH 8, 0.25 M LiCl, 0.5% NP-40, 0.5% sodium deoxycholate, 1 mM EDTA) and once in buffer F (10 mM Tris-HCl pH 8, 1 mM EDTA, 50 mM NaCl). Finally, input and immunoprecipitated chromatin samples were resuspended in a solution containing TE buffer/1% SDS, de-crosslinked by heating at 65 °C overnight and subjected to both RNase A (Invitrogen, 12091-039, 1 mg $ml^{-1}$) and Proteinase K (Thermo Scientific, EO0491, 20 mg $ml^{-1}$) treatments. Input and immunoprecipitated DNA extraction: after reverse crosslinking, input and immunoprecipitated chromatin samples were

treated with RNase A and proteinase K, and glycogen (Thermo Scientific, R0561, 20 mg ml$^{-1}$) was added. Samples were incubated at 37 °C for 2 h. DNA precipitation was carried out using 8 M LiCl (final concentration 0.44 M) and phenol:chloroform:isoamyl alcohol. Samples were vortexed and centrifuged at 20,000*g* for 15 min at 4 °C. The upper phase was mixed with chloroform by vortexing. After centrifugation at 20,000*g* for 15 min at 4 °C, the upper phase was mixed with –20 °C absolute ethanol by vortexing and stored at –80 °C for 2 h. Next, samples were pelleted at 20,000*g* for 20 min at 4 °C. The pellets were washed with ice-cold 70% ethanol and centrifuged at 20,000*g* for 15 min at 4 °C. The supernatants were discarded and pellets were dried at room temperature, dissolved in nuclease-free water and quantified using a Qubit fluorometric assay (Invitrogen) according to the manufacturer's protocol. Library preparation and sequencing: Illumina compatible libraries were prepared from input and immunoprecipitated DNAs using the Illumina TruSeq ChIP library preparation kit according to the manufacturer's protocol (IP-202-1012). In brief, 4 to 10 nanograms of DNA were subjected to end-repair, dA-tailing and ligation of TruSeq indexed Illumina adapters. After a final PCR amplification step (with 15 cycles), the resulting barcoded libraries were equimolarly pooled and quantified by quantitative PCR using the KAPA library quantification kit (Roche, 07960336001). Sequencing was performed on the NovaSeq 6000 (Illumina), targeting 75 million clusters per sample and using paired-end 2× 100 bp. ChIP–seq data processing and quality controls have been performed with the Institut Curie ChIP–seq Nextflow pipeline (1.0.6) available at https://github.com/bioinfo-pf-curie/ChIP-seq. In brief, reads were trimmed for adapter content, and aligned on the Human reference genome hg38 with BWA-mem. Low-quality mapped reads, reads aligned on ENCODE blacklist regions, reads aligned on the spike-in genome and reads marked as duplicates were discarded from the analysis. Bigwig tracks were then generated with deeptools and normalized to 1 million reads to account for differences in sequencing depth. In order to integrate the histone mark enrichments with the gene expression data (RNA-seq), the ChIP–seq signal has been counted either at the transcription start site level (±2 kb) for permissive histone marks or at the gene body for repressive histone marks. Coding genes from Gencode v34 have been used for the annotations. ChIP–seq counts data have then been filtered to remove low counts, and

normalized using the TMM methods (edgeR R package). Fold-changes have been then calculated for all genes and donors.

## In vivo animal studies

Survival assessment using the LPS mouse model was conducted at Fidelta (now Selvita) according to 2010/63/EU and National legislation regulating the use of laboratory animals in scientific research and for other purposes (Official Gazette 55/13). An institutional committee on animal research ethics (CARE-Zg) monitored animal-related procedures to ensure they were not compromising animal welfare. Experiments were performed on eight-week-old male BALB/c mice. Other experiments involving LPS and CLP mouse models were performed in accordance with French laws concerning animal experimentation (#2021072216346511) and approved by the Institutional Animal Care and Use Committee of Université de Saint-Quentin-en-Yvelines (C2EA-47). These LPS experiments were performed on eight-week-old male BALB/c mice and five-week-old male SWISS mice and experiments involving the CLP model were performed on nine-week-old male BALB/c mice. Mice were housed in a state-of-the-art animal care facility (2CARE, prefectural number agreement: A78-322-3, France). For experiments involving SARS-CoV-2, 8-week-old male K18-human ACE2-expressing C57BL/6 mice were used. A SARS-CoV-2 mouse model was used within the biosafety level 3 facility of the Institut Pasteur de Lille, after validation of the protocols by the local committee for the evaluation of the biological risks and complied with current national and institutional regulations and ethical guidelines (Institut Pasteur de Lille/B59-350009). The experimental protocols using animals were approved by the institutional ethical committee 'Comité d'Ethique en Experimentation Animale (CEEA) 75, Nord-Pas-de-Calais'. The animal study was authorized by the 'Education, Research and Innovation Ministry' under registration number APAFIS#25517-2020052608325772v3. A SARS-CoV-2 mouse model was used within the biosafety level 3 facility of the University of Toulouse. This work was overseen by an Institutional Committee on Animal Research Ethics (License APAFIS#27729-2020101616517580 v3, Minister of Research, France (CEEA-001)), to ensure that animal-related procedures were not compromising the animal welfare. Four mice per cage were housed in in a ventilated rack with a media enrichment element. Mice in all animal

facilities were housed in ventilated cages (temperature 22 °C ± 2 °C, humidity 55% ± 10%) with free access to water and food on a 12 h light/dark cycle. Male littermates were randomly assigned to experimental groups throughout.

## LPS-induced sepsis model

Survival assessment was conducted at Fidelta (now Selvita). LPS (Sigma-Aldrich, L2630, 20 mg kg$^{-1}$) was injected intraperitoneally to male BALB/c mice (8 weeks old). LCC-12 (0.3 mg kg$^{-1}$, intraperitoneal injection, $n = 10$) or vehicle (0.9% NaCl, 10 ml kg$^{-1}$, intraperitoneal injection, $n = 10$) were injected 2 h prior LPS challenge, then 24 h, 48 h, 72 h and 96 h after challenge. Dexamethasone (10 mg kg$^{-1}$, oral gavage, $n = 10$) was given 1 h before LPS challenge. Incidence of mortality was monitored every 4 h up to 48 h, then twice daily. The cytometry work, ICP-MS, western blotting and RNA-seq work were performed on eight-week-old male BALB/c mice and five-week-old male SWISS mice. The experimental endotoxemic model was induced by intraperitoneal injection of LPS (5 mg kg$^{-1}$ in BALB/c mice (*Escherichia coli* O111:B4, Sigma-Aldrich, L2630) or 20 mg kg$^{-1}$ in SWISS mice (*E. coli* O55:B5, Sigma-Aldrich, L2880)). All the mice were resuscitated with 30 ml kg$^{-1}$ body weight of saline administered subcutaneously 6 h after LPS administration. LCC-12 (0.3 mg kg$^{-1}$) was injected intraperitoneally. Mice were killed 22 h post LPS challenge. Body temperature was measured at 0 h, 6 h and 22 h post LPS challenge. Flow cytometry: after euthanasia at 22 h post LPS challenge, the organs were perfused with PBS/EDTA (1 ml g$^{-1}$, 2 mM, pH 7.4) and 10 ml of 1× PBS were injected in the peritoneum. The peritoneal liquid was then collected and centrifugated for 5 min at 1,500 rpm. The pellet was resuspended in RPMI medium containing 2% fetal calf serum (Dutscher, S181H-100). The peritoneal liquid was then washed in 96-well plates at 2,000 rpm for 2 min, and pellets were suspended in complete RPMI. For intracellular protein staining, samples were incubated for 2 h in brefeldin A (BFA, 5 ng ml$^{-1}$, Invitrogen, 00-4506-51) before surface staining with Fixable Viability Dye eFluor 780 (Invitrogen, 65-0865-14) followed by fluorochrome-conjugated antibodies (35 min at 4 °C). Samples incubated with BFA were fixed (Foxp3/Transcription factor Fix/Perm 4X, Cell Signaling, 44931S) for 20

min at 4 °C and permeabilized (Flow Cytometry Perm Buffer 10X, TONBO, TNB-1213-L150) before intracellular staining. The antibodies used were as follows: CD11b-Pacific Blue (BioLegend, 101224), CD40-APC (BioLegend, 124612), CD45-BV510 (BioLegend, 103138), CD86-PE (BioLegend, 105007), CD170 (Siglec-F)-PEeFluor 610 (eBioscience, 61-1702-80), F4/80-BV605 (BioLegend, 123133), I-A/I-E-AF700 (BioLegend, 107622), Ly6C–PerCP/Cy5.5 (BioLegend, 128012) and Ly6G–PE/CY7 (BioLegend, 127618). For intracellular staining, NOS2–APC (eBioscience, 17-5920-82) was used. SPMs correspond to $CD45^{+}$IA-$IE^{+}CD11b^{+}F4/80^{int}SiglecF^{-}$ cells. After washing with PBS or with Perm Buffer, data were acquired using an LSR Fortessa flow cytometer (BD Biosciences) and analysed with FlowJo. software v. 10.8.2. ICP-MS experiments were conducted using SPMs as described in 'ICP-MS'. Tissue-specific data were normalized against dry weight. Sorting of SPMs was performed using the following antibodies: CD11b-Pacific Blue (BioLegend, 101224), F4/80-PE (TONBO, TNB50-4801-U100), Ly6C–PerCP/Cy5.5 (BioLegend, 128012) and Ly6G–AF647 (BioLegend, 127610). The sorted SPMs corresponded to $CD11b^{+}F4/80^{int}Ly6C^{-}Ly6G^{-}$ cells and were isolated on a BD FACSAria IIIu. RNA-seq: the $CD11b^{+}F4/80^{int}Ly6C^{-}Ly6G^{-}$ sorted SPMs from 7 LPS-treated mice (102283 SPM) and from 8 LPS and LCC-12-treated SWISS mice (66967 SPM) were centrifuged, then resuspended in 350 µl of TCL (Qiagen, 1031576) and 1% β-mercaptoethanol. Total RNA was extracted with the Norgen Single Cell RNA Purification kit (Norgen, 51800). RNA sequencing libraries were prepared using the SMARTer Stranded Total RNA-seq Kit v2—Pico Input Mammalian (Clontech/Takara, 634419). The input quantity of total RNA was 10 ng for each condition. A first step of RNA fragmentation, was applied using a proprietary fragmentation mix at 94 °C for 4 min. After fragmentation, indexed cDNA synthesis was performed. Then the ribodepletion step was performed using probes specific to mammalian rRNA. PCR amplification (12 cycles) was done to amplify the cDNA libraries. Library quantification and quality assessment was performed using a Qubit fluorometric assay (Invitrogen, Q32854) with dsDNA HS (High Sensitivity) Assay Kit and LabChip GX Touch using a High Sensitivity DNA chip (Perkin Elmer, 760517, CLS760672). Libraries were then equimolarly pooled and quantified by quantitative PCR using the KAPA library quantification kit (Roche, 07960336001). Sequencing was performed on a NovaSeq 6000 (Illumina)

targeting 100 million clusters per sample and using paired-end 2× 100 bp. For the analysis, see 'RNA-seq'. Trimmed reads were aligned on the mouse mm10 reference genome.

## CLP-induced sepsis model

Nine-week-old male BALB/c mice were used for these experiments. Animals were anaesthetized by isoflurane (Forene). After abdominal incision, the cecum was ligated, punctured with a gauge needle (25G), and a small amount of fecal matter was released. For the sham group, after abdominal incision, the cecum was manipulated but was neither ligated nor punctured. After the cecum was returned to the abdomen, the abdominal cavity was closed in two layers and the mice were resuscitated with 30 ml $kg^{-1}$ body weight of saline (0.9% NaCl) administered subcutaneously. LCC-12 (0.3 mg $kg^{-1}$, intraperitoneal injection) was administered at 4 h, 24 h, 48 h, 72 h and 96 h following CLP creation. Mortality incidence was monitored every 2 h up to 120 h (except from 10 pm to 6 am) after CLP creation. Dexamethasone was administered intraperitoneally at 1 mg $kg^{-1}$ 5 min before CLP creation. ICP-MS experiments were conducted on SPMs as described in 'ICP-MS'. Tissue-specific data were normalized against dry weight. The sorting of SPMs was done using the following antibodies: CD11b-Pacific Blue (BioLegend, 101224), F4/80-PE (TONBO, TNB50-4801-U100), Ly6C–PerCP/Cy5.5 (BioLegend, 128012) and Ly6G–AF647 (BioLegend, 127610). The sorted SPMs corresponded to $CD11b^{+}F4/80^{int}Ly6C^{-}Ly6G^{-}$ cells.

## SARS-CoV-2-induced acute inflammation model

Eight-week-old male K18-human ACE2-expressing C57BL/6 mice (B6.Cg-Tg(K18-hACE2)2Prlmn/J) were purchased from Jackson Laboratory. The mice were anaesthetized by intraperitoneal injection of ketamine (100 mg $kg^{-1}$) and xylazine (10 mg $kg^{-1}$) and then intranasally infected with 50 µl of DMEM containing $5 \times 10^{2}$ $TCID_{50}$ of hCoV-19_IPL_France strain of SARS-CoV-2 (NCBI MW575140). LCC-12 was inoculated intranasally (0.5 mg $ml^{-1}$, 50 µl) 6 h, 24 h and 48 h post-infection. Mice were killed at day 4 post-infection. Cell sorting: K18-hACE2 mice (Jackson Laboratory, male, 10

mice per group) were infected under a short anaesthesia (isoflurane 4%), by the intranasal route with $10^5$ PFU of SARS-CoV-2 (strain Be-taCoV/France/IDF0372/2020). During the 4 days of infection, mice were monitored daily for disease progression. At day 4 post-infection, after a terminal anaesthesia (ketamine 100 mg $kg^{-1}$ and xylazine 10 mg $kg^{-1}$, intraperitoneal injection), lung tissues were collected and homogenized in gentleMACS C Tubes (Miltenyi Biotec) containing 2.5 ml of RPMI 1640 medium (Gibco) and collagenase (2.5 mg $ml^{-1}$, Roche) using a gentleMACS Dissociator (Miltenyi Biotec). Lung tissues were further dissociated for 30 min at 37 °C under shaking, passed through a 70 µm cell strainer, and proceeded for red blood cell lysis before staining for cell sorting. Cells were stained with Fixable Viability Dye eFluor 780 (Invitrogen, 65-0865-14) followed by fluorochrome-conjugated antibodies (35 min at 4 °C). The antibodies used were as follows: CD11b-Pacific Blue (BioLegend, 101224), CD45-BV510 (BioLegend, 103138), CD170 (Siglec-F)-PEeFluor 610 (eBioscience, 61-1702-80), F4/80-BV605 (BioLegend, 123133). AMs correspond to $CD45^{+}CD11b^{int}F4/80^{+}SiglecF^{+}$ cells and were isolated on a BD FACSAria Fusion. ICP-MS experiments were conducted in AMs as described in ‘ICP-MS’. RNA-seq: half of the right lobe was homogenized in 1 ml of RA1 buffer from the NucleoSpin RNA kit (Macherey Nagel, 740955.250) containing 20 mM of tris(2-carboxyethyl)phosphine (TCEP). Total RNAs in the tissue homogenate were extracted with the NucleoSpin RNA kit. RNA was eluted with 60 µl of water. For the following steps, see ‘RNA-seq’. Trimmed reads were aligned on the mouse mm10 reference genome.

## Software for illustrations

Illustrations were created using FIJI 2.0.0-rc-69/1.52n, Prism 8.2.0 and Adobe Illustrator 26.0.2 and BioRender.com. BioRender.com was used for Figs. 1a and 5a and Extended Data Fig. 10g.

## Quantification, statistical analysis and reproducibility

Results are presented as mean ± s.e.m or mean ± s.d. as indicated. In box plots, boxes represent interquartile range and median, and whiskers indicate the minimum and maximum values. A specific colour of dots on a box plot

represents a distinct donor only within a given figure panel. Each donor or mouse represents an independent biological sample. Prism 8.2.0 software was used to calculate *P* values using a two-sided Mann–Whitney test, two-sided unpaired *t*-test, Kruskal–Wallis test with Dunn's post test, two-way ANOVA or Mantel–Cox log-rank test as indicated. Prism 8.2.0 software or the R programming language was used to generate graphical representations of quantitative data unless stated otherwise. Exact *P* values are indicated in the figures. Sample sizes ($n$) are indicated in the figure legends. All immunofluorescence experiments were repeated with at least $n = 3$ donors with similar results. Western blotting on macrophages isolated from mice was performed on pooled samples and performed once per pool. Morphological changes observed between naMDMs and aMDMs were observed in $n = 128$ donors and representative images of $n = 1$ donor are displayed in Extended Data Fig. 1c. NanoSIMS imaging was performed on $n = 1$ donor and a representative image is displayed in Extended Data Fig. 5b.

### Materials availability

In-house reagents can be made available under a material transfer agreement with Institut Curie. Inquiries should be addressed to R.R.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

RNA-seq and ChIP–seq data are available at the Gene Expression Omnibus with accession reference GSE160864. The mass spectrometry proteomics raw data have been deposited to the ProteomeXchange Consortium via the PRIDE[80] partner repository with the dataset identifier PXD038612. The donor number corresponds to the order of blood collection. Source data are provided with this paper.

## Code availability

Analysis scripts for RNA-seq and ChIP–seq data are available at https://github.com/bioinfo-pf-curie/MDMmetals.

## References


1. Lopez-Otin, C. & Kroemer, G. Hallmarks of health. *Cell* **184**, 1929–1939 (2021).

2. Annane, D., Bellissant, E. & Cavaillon, J. M. Septic shock. *Lancet* **365**, 63–78 (2005).

3. Moore, J. B. & June, C. H. Cytokine release syndrome in severe COVID-19. *Science* **368**, 473–474 (2020).

4. Park, M. D., Silvin, A., Ginhoux, F. & Merad, M. Macrophages in health and disease. *Cell* **185**, 4259–4279 (2022).

5. Horby, P. et al. Dexamethasone in hospitalized patients with Covid-19. *N. Engl. J. Med.* **384**, 693–704 (2021).

6. Annane, D. et al. Hydrocortisone plus fludrocortisone for adults with septic shock. *N. Engl. J. Med.* **378**, 809–818 (2018).

7. Rudd, K. E. et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet* **395**, 200–211 (2020).

8. Aruffo, A., Stamenkovic, I., Melnick, M., Underhill, C. B. & Seed, B. CD44 is the principal cell surface receptor for hyaluronate. *Cell* **61**, 1303–1313 (1990).

9. Hua, Q., Knudson, C. B. & Knudson, W. Internalization of hyaluronan by chondrocytes occurs via receptor-mediated endocytosis. *J. Cell Sci.* **106**, 365–375 (1993).

10. Toole, B. P. Hyaluronan: from extracellular glue to pericellular cue. *Nat. Rev. Cancer* **4**, 528–539 (2004).

11. Ponta, H., Sherman, L. & Herrlich, P. A. CD44: from adhesion molecules to signalling regulators. *Nat. Rev. Mol. Cell Biol.* **4**, 33–45 (2003).

12. Brabletz, T., Kalluri, R., Nieto, M. A. & Weinberg, R. A. EMT in cancer. *Nat. Rev. Cancer* **18**, 128–134 (2018).

13. Guilliams, M. & Svedberg, F. R. Does tissue imprinting restrict macrophage plasticity? *Nat. Immunol.* **22**, 118–127 (2021).

14. Puré, E. & Cuff, C. A. A crucial role for CD44 in inflammation. *Trends Mol. Med.* **7**, 213–221 (2001).

15. Teder, P. et al. Resolution of lung inflammation by CD44. *Science* **296**, 155–158 (2002).

16. Bartolazzi, A., Peach, R., Aruffo, A. & Stamenkovic, I. Interaction between CD44 and hyaluronate is directly implicated in the regulation of tumor development. *J. Exp. Med.* **180**, 53–66 (1994).

17. Zoltan-Jones, A., Huang, L., Ghatak, S. & Toole, B. P. Elevated hyaluronan production induces mesenchymal and transformed properties in epithelial cells. *J. Biol. Chem.* **278**, 45801–45810 (2003).

18. Zöller, M. CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? *Nat. Rev. Cancer* **11**, 254–267 (2011).

19. Müller, S. et al. CD44 regulates epigenetic plasticity by mediating iron endocytosis. *Nat. Chem.* **12**, 929–938 (2020).

20. McKee, C. M. et al. Hyaluronan (HA) fragments induce chemokine gene expression in alveolar macrophages. The role of HA size and CD44. *J. Clin. Invest.* **98**, 2403–1243 (1996).

21. Kruidenier, L. et al. A selective jumonji H3K27 demethylase inhibitor modulates the proinflammatory macrophage response. *Nature* **488**, 404–408 (2012).

22. Saeed, S. et al. Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* **345**, 1251086 (2014).

23. Netea, M. G. et al. Trained immunity: A program of innate immune memory in health and disease. *Science* **352**, aaf1098 (2016).

24. Menke-van der Houven van Oordt, C. W. et al. First-in-human phase I clinical trial of RG7356, an anti-CD44 humanized antibody, in patients with advanced, CD44-expressing solid tumors. *Oncotarget* **7**, 80046–80058 (2016).

25. Madau, M. et al. A mild and straightforward one-pot hyaluronic acid functionalization through termination of poly-(2-alkyl-2-oxazoline). *Polymer* **230**, 124059 (2021).

26. Ren, M., Deng, B., Wang, J. Y., Liu, Z. R. & Lin, W. A dual-emission fluorescence-enhanced probe for imaging copper(ii) ions in lysosomes. *J. Mater. Chem. B* **3**, 6746–6752 (2015).

27. Slotta, K. H. & Tschesche, R. Über Biguanide, I.: Zur Konstitution der Schwermetall-Komplexverbindungen des Biguanids. *Ber. dt. chem. Ges.* **62**, 1390–1398 (1929).

28. Zhu, M., Lu, L., Yang, P. & Jin, X. Bis(1,1-dimethylbiguanido)copper(ii) octahydrate. *Acta. Cryst.* **E58**, m217–m219 (2002).

29. Zhou, G. et al. Role of AMP-activated protein kinase in mechanism of metformin action. *J. Clin. Invest.* **108**, 1167–1174 (2001).

30. Ge, E. J. et al. Connecting copper and cancer: from transition metal signalling to metalloplasia. *Nat. Rev. Cancer* **22**, 102–113 (2022).

31. Wang, J. et al. Inhibition of human copper trafficking by a small molecule significantly attenuates cancer cell proliferation. *Nat. Chem.* **7**, 968–979 (2015).

32. Cui, L. et al. Mitochondrial copper depletion suppresses triple-negative breast cancer in mice. *Nat. Biotechnol.* **39**, 357–367 (2021).

33. Rodriguez, R., Schreiber, S. L. & Conrad, M. Persister cancer cells: Iron addiction and vulnerability to ferroptosis. *Mol. Cell* **82**, 728–740 (2022).

34. Cañeque, T., Müller, S. & Rodriguez, R. Visualizing biologically active small molecules in cells using click chemistry. *Nat. Rev. Chem.* **2**, 202–215 (2018).

35. Tornoe, C. W., Christensen, C. & Meldal, M. Peptidotriazoles on solid phase: [1,2,3]-triazoles by regiospecific copper(i)-catalyzed 1,3-dipolar cycloadditions of terminal alkynes to azides. *J. Org. Chem.* **67**, 3057–3064 (2002).

36. Rostovtsev, V. V., Green, L. G., Fokin, V. V. & Sharpless, K. B. A stepwise huisgen cycloaddition process: copper(I)-catalyzed regioselective "ligation" of azides and terminal alkynes. *Angew. Chem. Int. Ed.* **41**, 2596–2599 (2002).

37. Sletten, E. M. & Bertozzi, C. R. Bioorthogonal chemistry: fishing for selectivity in a sea of functionality. *Angew. Chem. Int. Ed.* **48**, 6974–6998 (2009).

38. Tsvetkov, P. et al. Copper induces cell death by targeting lipoylated TCA cycle proteins. *Science* **375**, 1254–1261 (2022).

39. Wang, L. et al. Fluorescence imaging mitochondrial copper(ii) via photocontrollable fluorogenic probe in live cells. *Chinese Chem. Lett.* **28**, 1965–1968 (2017).

40. Chan, P. C. & Kesner, L. Copper (ii) complex-catalyzed oxidation of NADH by hydrogen peroxide. *Biol. Trace Elem. Res.* **2**, 159–174 (1980).

41. Robbins, M. H. & Drago, R. S. Activation of hydrogen peroxide for oxidation by copper(ii) complexes. *J. Cat.* **170**, 295–303 (1997).

42. Dai, Z., Ramesh, V. & Locasale, J. W. The evolving metabolic landscape of chromatin biology and epigenetics. *Nat. Rev. Genet.* **21**, 737–753 (2020).

43. Chua, R. L. et al. COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* **38**, 970–979 (2020).

44. Pai, A. A. et al. Widespread shortening of 3′ untranslated regions and increased exon inclusion are evolutionarily conserved features of innate immune responses to infection. *PLoS Genet.* **12**, e1006338 (2016).

45. Fernandes, M. C. et al. Dual transcriptome profiling of Leishmania-infected human macrophages reveals distinct reprogramming

signatures. *mBio* **7**, e00027-16 (2016).

46. Gonçalves, S. M. et al. Phagosomal removal of fungal melanin reprograms macrophage metabolism to promote antifungal immunity. *Nat. Commun.* **11**, 2282 (2020).

47. Hoffmann, J. A., Kafatos, F. C., Janeway, C. A. & Ezekowitz, R. A. Phylogenetic perspectives in innate immunity. *Science* **284**, 1313–1318 (1999).

48. Buras, J. A., Holzmann, B. & Sitkovsky, M. Animal models of sepsis: setting the stage. *Nat. Rev. Drug Discov.* **4**, 854–865 (2005).

49. Kulkarni, A. S., Gubbi, S. & Barzilai, N. Benefits of metformin in attenuating the hallmarks of aging. *Cell Metab.* **32**, 15–30 (2020).

50. Bharath, L. P. et al. Metformin enhances autophagy and normalizes mitochondrial function to alleviate aging-associated inflammation. *Cell Metab.* **32**, 44–55.e6 (2020).

51. Messaoudii, C., Boudier, T., Sanchez Sorzano, C. O. & Marco, S. TomoJ: tomography software for three-dimensional reconstruction in transmission electron microscopy. *BMC Bioinf.* **8**, 288–296 (2007).

52. Lechene, C. et al. High-resolution quantitative imaging of mammalian and bacterial cells using stable isotope mass spectrometry. *J. Biol.* **5**, 20 (2006).

53. Gräber, M. et al. Oral disinfectants inhibit protein-protein interactions mediated by the anti-apoptotic protein Bcl-xL and induce apoptosis in human oral tumor cells. *Angew. Chem. Int. Ed. Engl.* **52**, 4487–4491 (2013).

54. Giguère, J.-B. et al. Synthesis of [2]-and [3] rotaxanes through Sonogashira coupling. *Tetrahedron Lett.* **50**, 5497–5500 (2009).

55. Valle, F. D. & Romeo, A. Esters of hyaluronic acid. US patent US-4851521-A (1986).

56. Zhu, M., Lu, L., Yang, P. & Jin, X. Bis(1,1-dimethylbiguanido)copper(ii) octahydrate. *Acta Crystallogr. E* **58**, m217–m219 (2002).

57. Allouche, A.-R. Gabedit-A graphical user interface for computational chemistry software. *J. Comput. Chem.* **32**, 174–182 (2011).

58. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

59. Stewart, J. J. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **19**, 1–32 (2013).

60. Klamt, A. & Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin. Trans.* **2**, 799–805 (1993).

61. Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **8**, e1327 (2018).

62. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).

63. Zhao, Y. & Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **120**, 215–241 (2007).

64. Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *J. Chem. Phys.* **98**, 1372–1377 (1993).

65. Lee, C., Yang, W. & Parr, R. G. Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).

66. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297 (2005).

67. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).

68. Galván-García, E. A., Agacino-Valdés, E., Franco-Pérez, M. & Gómez-Balderas, R. $[Cu(H_2O)_n]^{2+}$ ($n = 1–6$) complexes in solution phase: a DFT hierarchical study. *Theor. Chem. Acc.* **136**, 29 (2017).

69. Schäfer, A., Horn, H. & Ahlrichs, R. Fully optimized contracted gaussian basis sets for atoms Li to Kr. *J. Chem. Phys.* **97**, 2571–2577 (1992).

70. Eichkorn, K., Weigend, F., Treutler, O. & Ahlrichs, R. Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials. *Theor. Chem. Acc.* **97**, 119–124 (1997).

71. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **113**, 6378–6396 (2009).

72. Ribeiro, R. F., Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Use of solution-phase vibrational frequencies in continuum models for the free energy of solvation. *J. Phys. Chem. B* **115**, 14556–14562 (2011).

73. Viltard, M. et al. The metabolomic signature of extreme longevity: naked mole rats versus mice. *Aging* **11**, 4783–4800 (2019).

74. Poullet, P., Carpentier, S. & Barillot, E. myProMS, a web server for management and validation of mass spectrometry-based proteomic data. *Proteomics* **7**, 2553–2556 (2007).

75. The, M., MacCoss, M. J., Noble, W. S. & Kall, L. Fast and accurate protein false discovery rates on large-scale proteomics data sets with Percolator 3.0. *J. Am. Soc. Mass. Spectrom.* **27**, 1719–1727 (2016).

76. Valot, B., Langella, O., Nano, E. & Zivy, M. MassChroQ: a versatile tool for mass spectrometry quantification. *Proteomics* **11**, 3572–3577 (2011).

77. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).

78. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

79. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

80. Perez-Riverol, Y. et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).

## Acknowledgements

R.R. thanks J.-M. Lehn, S. Schreiber and R. Vale. This work was supported by the CNRS, INSERM, PSL Research University, Paris Saclay University, Ecole Polytechnique, University of Bath, University of Lille and Institut Pasteur (Lille). R.R. was supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 647973), Foundation Charles Defforey-Institut de France, Region IdF for NMR infrastructure, Ligue Contre le Cancer and the Ladies of Pompadour. G.K. was supported by ANR Grant Wilsonmed and LabEx Immuno-Oncology (ANR-18-IDEX-0001). D.A. was supported by FHU Sepsis and Programme d'Investissements d'avenir ANR-18-RHUS-0004-RHU records. ICP-MS platform at Institut de Physique du Globe de Paris is supported by IPGP multidisciplinary

programme PARI and Region IdF (SESAME grant agreement no. 12015908). This work was granted access to the HPC resources of CINES under the allocation 2020-A0070810977 made by GENCI, the Balena High Performance Computing (HPC) Service at the University of Bath, the PICT-IbiSA@BDD Imaging Facility of Institut Curie, member of the France-BioImaging national research infrastructure (ANR-10-INBS-04), the ICGex NGS platform of Institut Curie (ANR-10-INBS-09-08, INCa-DGOS-465, INCa-DGOS-Inserm_12554), GenoToul ANEXPLO Animal Level 3 and TRI Facilities of the IPBS (Investissement d'Avenir and Foundation Bettencourt), the mass spectrometry platform supported by Région Ile-de-France (N°EX061034) and ITMO Cancer of Aviesan and INCa on funds administered by INSERM (N°21CQ016-00), and the cytometry platform of Institut Curie. We thank P. Benaroch, J.-L. Guerquin-Kern, M. Plays, J. Sampaio Lopes, D. Guillemot, O. Delattre and O. Neyrolles for support.

# Author information

Author notes

1. These authors contributed equally: Stéphanie Solier, Sebastian Müller, Tatiana Cañeque, Antoine Versini

## Authors and Affiliations

1. Equipe Labellisée Ligue Contre le Cancer, Institut Curie, CNRS, INSERM, PSL Research University, Paris, France

   Stéphanie Solier, Sebastian Müller, Tatiana Cañeque, Antoine Versini, Fabien Sindikubwabo, Leeroy Baron, Christine Gaillet, Alain Puisieux & Raphaël Rodriguez

2. Paris Saclay University, UVSQ, INSERM, 2I, Montigny-le-Bretonneux, France

   Arnaud Mansart, Laila Emam & Djillali Annane

3. CBIO-Centre for Computational Biology, Institut Curie, INSERM, Mines ParisTech, Paris, France

Pierre Gestraud & Nicolas Servant

4. Department of Chemistry, University of Bath, Bath, UK

G. Dan Pantoș

5. Institut de Chimie Moléculaire et des Matériaux d’Orsay, CNRS, Paris Saclay University, Orsay, France

Vincent Gandon

6. Laboratoire de Chimie Moléculaire, CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, Palaiseau, France

Vincent Gandon

7. Institut Curie, PSL Research University, Paris, France

Ting-Di Wu

8. Multimodal Imaging Center, Paris Saclay University, CNRS, INSERM, Orsay, France

Ting-Di Wu

9. CurieCoreTech Mass Spectrometry Proteomic, Institut Curie, PSL Research University, Paris, France

Florent Dingli & Damarys Loew

10. ICGex Next-Generation Sequencing Platform, Institut Curie, PSL Research University, Paris, France

Sylvain Baulande

11. Metabolomics and Cell Biology Platforms, Institut Gustave Roussy, Villejuif, France

Sylvère Durand & Guido Kroemer

12. Université de Lille, CNRS, INSERM, CHU Lille, Institut Pasteur de Lille, CIIL, Lille, France

Valentin Sencio, Cyril Robil & François Trottein

13. Institut of Pharmacology and Structural Biology, University of Toulouse, CNRS, Toulouse, France

David Péricat, Emmanuelle Näser, Céline Cougoule & Etienne Meunier

14. Cytometry and Imaging Core facility, Institute of Pharmacology and Structural Biology, University of Toulouse, CNRS, Toulouse, France

Emmanuelle Näser

15. Institut Curie, INSERM, PSL Research University, Paris, France

Anne-Laure Bègue, Hélène Salmon & Nicolas Manel

16. Department of Medical Oncology, Institut Curie, PSL Research University, Paris, France

Sarah Watson

17. Peter MacCallum Cancer Centre and Sir Peter MacCallum Department of Oncology, Melbourne, Victoria, Australia

Mark A. Dawson

18. Centre for Cancer Research, University of Melbourne, Melbourne, Victoria, Australia

Mark A. Dawson

19. Centre de Recherche des Cordeliers, University of Paris, Sorbonne University, INSERM, Institut Universitaire de France, Paris, France

Guido Kroemer

20. Institut du Cancer Paris CARPEM, Department of Biology, Hôpital Européen Georges Pompidou, AP-HP, Paris, France

Guido Kroemer

21. Department of Intensive Care, Hôpital Raymond Poincaré, AP-HP, Garches, France

Djillali Annane

### Contributions

R.R. conceptualized the study and directed the research. R.R., S.M., S.S. and T.C. designed the experiments and analysed the data. S.M. and S.S. performed the experiments unless stated otherwise. T.C., A.V., L.B. and C.G. performed chemical synthesis and analytical chemistry. G.D.P. and V.G. assisted with molecular modelling. T.-D.W. assisted with NanoSIMS imaging. S.D. and G.K. assisted with mass spectrometry-based metabolomics. F.S., F.D. and D.L. performed mass spectrometry-based proteomics. F.S., P.G., S.B. and N.S. performed Solexa/Illumina sequencing and bioinformatics. A.M., L.E. and D.A. established the preclinical models of sepsis and performed SPM isolation. V.S., C.R., F.T., D.P., E.N., C.C. and E.M. performed SARS-CoV-2 infection and alveolar macrophage isolation. R.R. and S.M. wrote the article. All authors contributed to drafts of the article.

### Corresponding author

Correspondence to Raphaël Rodriguez.

## Ethics declarations

### Competing interests

Institut Curie and the CNRS have filed patents on the LCC family of compounds and their therapeutic use. Patents: WO 2019/233982, filed on 4 June 2019; PCT/EP2021/082073, filed on 18 November 2021, WO 2021/233962, filed on 19 May 2021.

## Peer review

### Peer review information

*Nature* thanks Donita Brady and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

## Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Extended data figures and tables

### Extended Data Fig. 1 CD44 mediates the uptake of metals in inflammatory macrophages.

**a**, Flow cytometry of cell surface markers in MDM. Monocytes treated as indicated to obtain pro-inflammatory or anti-inflammatory states. **b**, Flow cytometry of cell surface markers in MDM. Data representative of $n = 13$ donors. **c**, Bright field microscopy images of MDM. Scale bar, 20 µm. Morphological observations representative of $n = 128$ donors. **d**, ICP-MS of cellular metals in MDM ($n = 9$ donors). **e**, ICP-MS of cellular metals in aMDM under knockdown conditions of indicated genes ($n = 6$ donors). **f**, Western blots of cellular metal transporters in aMDM under knockdown conditions of indicated genes ($n = 4$ donors). **g**, Western blots of cellular metal transporters in aMDM under CD44 knockdown conditions (*n*

= 6 donors). For **d**, **f** and **g** two-sided Mann-Whitney test. For **e**, Kruskal-Wallis test with Dunn's post-test. Box plots: boxes represent interquartile range and median and whiskers indicate the minimum and maximum values. Each colored dot represents a distinct donor for a given panel.

Source data

## Extended Data Fig. 2 CD44 mediates the uptake of metals in inflammatory macrophages.

**a**, ICP-MS of cellular metals in aMDM treated with an anti-CD44 antibody RG7356 during activation (*n* = 7 donors). **b**, ICP-MS of cellular metals in aMDM supplemented with HA (0.6-1 MDa) or permethylated (meth-) HA during activation (*n* = 6 donors). **c**, Western blots of hyaluronan synthases (HAS) (*n* = 6 donors) and ATP7A/B (*n* = 8 donors) in MDM. **d**, Fluorescence microscopy of a lysosomal copper(II) probe (Lys-Cu) in aMDM under CD44 knockdown conditions (*n* = 5 donors). **e**, Fluorescence microscopy of CTR2 and Lamp2 in aMDM (*n* = 4 donors). **f**, Fluorescence microscopy of a lysosomal copper(II) probe (Lys-Cu) in aMDM under CTR2 knockdown conditions (*n* = 5 donors). For **d**–**f**, Scale bars, 10 µm. For **a**, **c**, **d** and **f** two-sided Mann-Whitney test. For **b**, Kruskal-Wallis test with Dunn's post-test. Box plots: boxes represent interquartile range and median, and whiskers indicate the minimum and maximum values. Each colored dot represents a distinct donor for a given panel.

Source data

## Extended Data Fig. 3 Development of a small molecule inactivator of mitochondrial copper(II).

**a**, Flow cytometry of MDM treated with ATTM (10 µM, *n* = 5 donors), D-Pen (250 µM, *n* = 5 donors), EDTA (500 µM, *n* = 5 donors) or Trien (200 µM, *n* = 5 donors). **b**, Structural analysis of biguanide-based copper(II) complexes by molecular modeling. Top and side views highlight distinct geometries of $Cu(Met)_2$, Cu–LCC-12 and Cu–LCC-4,4. **c**, HRMS of $Cu(Met)_2$ and Cu–LCC-4,4. **d**, HRMS of LCC-12 in the presence of metals

as indicated. **e**, UV absorbance spectra of LCC-12 (5 µM) titrated with a solution of copper(II). **f**, Picture of aq. solutions of Met, LCC-12, $CuCl_2$ and corresponding mixtures. **g**, Western blots of AMPKα and phosphorylated AMPKα (p-AMPKα) in MDM treated with LCC-12 or Met ($n = 6$ donors). For **a** and **g** Kruskal-Wallis test with Dunn's post-test. Box plots: boxes represent interquartile range and median, and whiskers indicate the minimum and maximum values. Each colored dot represents a distinct donor for a given panel.

Source data

## Extended Data Fig. 4 Targeting mitochondrial copper(II) interferes with cell plasticity.

**a**, Flow cytometry of CD44 in dendritic cells (DC) ($n = 6$ donors), non-activated (naDC) and activated (aDC), $CD4^+$ T cells (CD4) ($n = 6$ donors), non-activated (naCD4) and activated (aCD4), $CD8^+$ T cells (CD8) ($n = 6$ donors), non-activated (naCD8) and activated (aCD8), CSF1/IL-4 anti-inflammatory macrophages ($n = 6$ donors), non-activated (naMDM2) and activated (aMDM2), neutrophils (N) ($n = 6$ donors), non-activated (naN) and activated (aN). **b**, Flow cytometry of cell surface markers in immune cells treated with LCC-12 during activation. **c**, Primary human non-small cell lung circulating cancer cells treated as indicated. Left: Flow cytometry of CD44. Middle: ICP-MS of cellular copper ($n = 5$ independent biological experiments). Right: Western blots of EMT markers. **d**, Murine pancreatic cancer cells treated as indicated. Left: Flow cytometry of CD44. Middle: ICP-MS of cellular copper ($n = 4$ independent biological experiments). Right: Western blots of EMT markers. For **b**–**d** two-sided Mann-Whitney test. Box plots: boxes represent interquartile range and median, and whiskers indicate the minimum and maximum values. Each colored dot represents a distinct donor for a given cell type.

Source data

## Extended Data Fig. 5 Detection of a druggable pool of copper(II) in mitochondria.

**a**, Molecular structure of isotopologue $^{15}N,^{13}C$-LCC-12. **b**, NanoSIMS image of $^{15}N$ and $^{197}Au$ in aMDM of *n* = 1 donor. **c**, Fluorescence microscopy of labelled LCC-12,4 (100 nM) in aMDM. In-cell-labelling performed without added copper(II) and using LCC-12 as a competitor. Representative of *n* = 3 donors. **d**, ICP-MS of metals in mitochondria of MDM (*n* = 6 donors). **e**, Comparison of the total metal contents in cells and mitochondria of MDM determined by ICP-MS. **f**, ICP-MS of metals in nuclei isolated from MDM (*n* = 6 donors). **g**, ICP-MS of metals in endoplasmic reticula (ER) isolated from MDM (*n* = 6 donors). **h**, ICP-MS of the total cellular copper content in MDM treated with LCC-12 (*n* = 6 donors). **i**, ICP-MS of mitochondrial copper in MDM treated with LCC-12 (*n* = 6 donors). **j**, Flow cytometry of a mitochondrial copper(II) probe ($M_{Cu}$-2) in MDM treated with LCC-12 (*n* = 10 donors). **k**, Western blots of mitochondrial metal transporters in MDM (*n* = 8 donors). **l**, **m**, **n**, Fluorescence microscopy of labelled LCC-12,4 in aMDM under gene knockdown conditions as indicated (*n* = 4 donors). **o**, Top: Structure of trientine alkyne. Bottom: Picture of aq. solutions of trientine alkyne, $CuSO_4$ and corresponding mixtures. **p**, Fluorescence microscopy of labelled trientine alkyne in aMDM. For **b**, **c**, **l** – **n**, **p** scale bar, 10 µm. For **c** two-sided unpaired t-test, representative of *n* = 3 donors. Mean ± s.d. For **d**, **f**, **g**, **k**, **l** – **n**, two-sided Mann-Whitney test. For **h** – **j** Kruskal-Wallis test with Dunn's post-test. Box plots: boxes represent interquartile range and median, and whiskers indicate the minimum and maximum values. Each colored dot represents a distinct donor for a given panel.

Source data

## Extended Data Fig. 6 Copper(II) regulates NAD(H) redox cycling.

**a**, $^{1}H$ NMR spectra of NADH (black) and the reaction products of NADH with $H_2O_2$ after 1 h at 37 °C, spectra recorded at 298 K in $D_2O$ (red). **b**, Reaction of MDHNA with $H_2O_2$ under copper(II)-catalyzed and copper-free conditions to afford either $MNA^+$ or a product of epoxidation, respectively. The mass of molecular ions detected by mass spectrometry are indicated. **c**, Heteronuclear single quantum coherence (HSQC) NMR

spectra of MDHNA and its product of epoxidation. Red stars mark $^{1}H$ and $^{13}C$ NMR signals of the most reactive double bond towards $H_2O_2$ and that of the corresponding epoxide product. Blue stars mark $^{1}H$ and $^{13}C$ NMR signals of the least reactive double bond towards $H_2O_2$. Blue boxes show $^{1}H$-$^{13}C$ HSQC correlations of the least reactive double bond. Red boxes show $^{1}H$-$^{13}C$ HSQC correlations of the most reactive double bond and corresponding epoxide. **d**, $^{1}H$ NMR spectra of NADH, imidazole and copper(II) (black), $NAD^+$ in the presence of imidazole and copper(II) (blue), the reaction product of NADH with $H_2O_2$ in the presence of imidazole and copper(II) after 1 h at 25 °C, spectra recorded at 298 K in buffered $D_2O$ (pD 8.4) (green). **e**, Free energy profile ($\Delta G_{298}$, kcal/mol) of the [(Imidazole)$_3$Cu($H_2O$)](II)-mediated H-transfer reaction from MDHNA to $H_2O_2$. Selected distances in Å. Free energy profile ($\Delta G_{298}$, kcal/mol) of the copper-free H-transfer reaction from MDHNA to $H_2O_2$. Selected distances in Å. **f**, Optimized Cu(Met)$_2$, MDHNA, $H_2O_2$ and $H_2O$. **g**, Concentrations of copper and NADH in cells, and in the cell-free system used in Fig. 3f. **h**, Quantification of extracellular lactate produced by MDM treated with LCC-12 or Met ($n = 5$ donors). **i**, Quantification of glyceraldehyde 3-phosphate (GA3P) in MDM treated with LCC-12 or Met ($n = 5$ donors). For **h** and **i** Kruskal-Wallis test with Dunn's post-test. Box plots: boxes represent interquartile range and median, and whiskers indicate the minimum and maximum values. Each colored dot represents a distinct donor for a given panel.

Source data

## Extended Data Fig. 7 Comparative analysis of transcriptomes of inflammatory macrophages.

**a**, Principal Component Analysis (PCA) of RNA-seq comparing naMDM ($n = 10$ donors) and aMDM ($n = 10$ donors) with MDM exposed to *Salmonella typhimurium* ($n = 32$ donors) vs control ($n = 32$ donors), macrophages from bronchoalveolar fluids of moderate ($n = 3$ donors) and severe COVID-19 individuals (sCOVID, $n = 6$ donors) vs control ($n = 4$ donors), MDM exposed to *Leishmania major* ($n = 5$ donors) vs control ($n =$

6 donors) and MDM exposed to *Aspergillus fumigatus* (*n* = 3 donors) vs control (*n* = 3 donors). **b**, GO term analyses of upregulated genes in MDM exposed to *Salmonella typhimurium* (*n* = 32 donors) vs control (*n* = 32 donors), sCOVID (*n* = 6 donors) vs control (*n* = 4 donors), *Leishmania major* (*n* = 5 donors) vs control (*n* = 6 donors) and *Aspergillus fumigatus* (*n* = 3 donors) vs control (*n* = 3 donors). **c**, RNA-seq analyses of gene expression in MDM exposed to *Salmonella typhimurium* (*n* = 32 donors) vs control (*n* = 32 donors), sCOVID (*n* = 6 donors) vs control (*n* = 4 donors), *Leishmania major* (*n* = 5 donors) vs control (*n* = 6 donors) and *Aspergillus fumigatus* (*n* = 3 donors) vs control (*n* = 3 donors). Inflammatory signature genes are highlighted. Dashed lines, adjusted *P* values = 0.05. **d**, RNA-seq analyses of gene expression in MDM exposed to *Salmonella typhimurium* (*n* = 32 donors) vs control (*n* = 32 donors), sCOVID (*n* = 6 donors) vs control (*n* = 4 donors), *Leishmania major* (*n* = 5 donors) vs control (*n* = 6 donors) and *Aspergillus fumigatus* (*n* = 3 donors) vs control (*n* = 3 donors). Genes encoding iron-dependent demethylases and acetyl-transferases are highlighted. Dashed lines, adjusted *P* values = 0.05. For **b** – **d** differential gene expression was assessed with the limma/voom framework. GO enrichment was assessed with the enrichGO method from clusterProfiler. *P* values were corrected for multiple testing with the Benjamini-Hochberg procedure.

## Extended Data Fig. 8 Mitochondrial copper(II) regulates epigenetic states and transcriptional programs of inflammatory macrophages.

**a**, Quantitative mass-spectrometry-based proteomics of MDM (*n* = 8 donors). **b**, Representative western blots (top) of epigenetic modifiers identified by RNA-seq in aMDM and corresponding quantifications (bottom) (*n* = 6–8 donors). **c**, Genes encoding iron-dependent demethylases and acetyl-transferases found to be upregulated in aMDM are listed together with putative substrates and post-translational modifications (PTMs) products. **d**, Fluorescence microscopy quantifications of histone H3 methyl and acetyl marks in MDM. Quantifications represent aMDM normalized against naMDM. At least 50 cells were quantified per donor per condition (*n* = 5–11 donors). **e**, Scatter plot correlation of a representative donor of ChIP-seq reads count of histone marks in genes against RNA-seq

of gene transcripts in MDM ($n = 10$ donors). **f**, ChIP-seq tracks of selected genes involved in inflammation in MDM. For **b** and **d**, two-sided Mann-Whitney test. Box plots: boxes represent interquartile range and median, and whiskers indicate the minimum and maximum values. Each colored dot represents a distinct donor for a given panel.

Source data

## Extended Data Fig. 9 Mitochondrial copper(II) regulates epigenetic states and transcriptional programs of inflammatory macrophages.

**a**, Representative western blots (top) of proteins involved in inflammation in MDM treated with LCC-12 and corresponding quantifications (bottom) ($n = 8$ donors). **b**, Immunoassay of cytokines secreted by MDM treated with LCC-12 during activation ($n = 6$ donors). **c**, PCA of RNA-seq comparing naMDM ($n = 10$ donors), aMDM ($n = 10$ donors) and MDM treated with LCC-12 during activation ($n = 5$ donors). **d**, Fluorescence microscopy quantifications of histone H3 methyl and acetyl marks in MDM treated with LCC-12. Quantifications were normalized against naMDM. At least 50 cells were quantified per donor per condition ($n = 5$–$7$ donors). **e**, Scatter plot correlation of a representative donor of ChIP-seq reads count of histone marks in genes against RNA-seq of gene transcripts in aMDM ($n = 10$ donors) and MDM treated with LCC-12 during activation ($n = 5$ donors). **f**, ChIP-seq tracks of selected genes involved in inflammation in MDM. **g**, Western blots of proteins involved in inflammation in aMDM under SOD2 knockdown conditions. **h**, Western blots of proteins involved in inflammation in aMDM under SLC25A3 knockdown conditions. **i**, Flow cytometry of wild-type (WT) and CD44 knockout (KO) aMDM. Gating strategy see Supplementary Information. **j**, Western blots of metal transporter proteins in WT and CD44-KO aMDM in $n = 1$ donor. **k**, Western blots of proteins involved in inflammation and histone marks in WT aMDM and CD44-KO aMDM for $n = 3$ donors. H3 is a sample processing control. For **a** and **d** Kruskal-Wallis test with Dunn's post-test. For **b**, **g** and **h** two-sided Mann-Whitney test. Box plots: boxes represent interquartile range and median, and whiskers indicate the minimum and maximum values. Each colored dot represents a distinct donor for a given panel.

Source data

## Extended Data Fig. 10 Pharmacological inactivation of mitochondrial copper(II) attenuates inflammation in vivo.

**a**, Western blots of copper-signalling effectors in SPMs from mice treated with LPS. Macrophages of several mice were pooled (4–7 mice per condition). **b**, Western blots of copper-signalling effectors in SPMs from mice subjected to CLP. Macrophages of several mice were pooled (7–8 mice per condition). H3 is a sample processing control. **c**, Western blots of copper-signalling effectors in AMs from K18-hACE2 mice infected with SARS-CoV-2. Macrophages of several mice were pooled (10 mice per condition). H3 is a sample processing control. **d**, Average body temperature of mice treated as indicated ($n = 6–9$ mice per group). **e**, GO term analysis of downregulated genes in lung tissues of SARS-CoV-2 infected K18-hACE2 mice treated with LCC-12 (0.5 mg/kg). **f**, RNA-seq analysis of gene expression in lung tissues of SARS-CoV-2-infected K18-hACE2 mice treated with LCC-12 (0.5 mg/kg) ($n = 8$ mice per group). Inflammatory signature genes highlighted. Dashed lines, adjusted $P$ value = 0.05. **g**, Illustration of copper-signalling. Cell plasticity involves upregulation of the cell surface marker CD44, which mediates endocytosis of metal-bound hyaluronates. In the presence of copper(II), NADH reacts with $H_2O_2$ to replenish $NAD^+$ in mitochondria, an enzyme cofactor involved in the biosynthesis of αKG and acetyl-CoA. These co-substrates of iron-dependent demethylases and acetyl-transferases are required for epigenetic and transcriptional programming of inflammation and the regulation of cell plasticity. Pharmacological inactivation of mitochondrial copper(II) blocks NAD(H) redox cycling, leading to distinct epigenetic states and transcriptional profiles. Targeting copper(II) interferes with cell plasticity in immune and cancer cells. For **a** – **c** gating strategy of SPMs and AMs see Methods and Supplementary Information. For **d** 2-way ANOVA. Mean values ± s.e.m. For **e** and **f** differential gene expression was assessed with the limma/voom framework. GO enrichment was assessed with the enrichGO method from clusterProfiler. $P$-values were corrected for multiple testing with the Benjamini-Hochberg procedure.

Source data

# Supplementary information

## Supplementary Information

This file contains Supplementary Table legends; sequences for RNA interference and genome-editing experiments; flow cytometry gating strategies; full western blots; and NMR spectra.

## Reporting Summary

## Peer Review File

## Supplementary Table 1

## Supplementary Table 2

## Supplementary Table 3

## Supplementary Table 4

## Supplementary Table 5

## Supplementary Table 6

## Supplementary Table 7

## Supplementary Table 8

## Supplementary Table 9

## Supplementary Table 10

## Supplementary Table 11

Supplementary Table 12

## Source data

Source Data Fig. 1

Source Data Fig. 2

Source Data Fig. 3

Source Data Fig. 5

Source Data Extended Data Fig. 1

Source Data Extended Data Fig. 2

Source Data Extended Data Fig. 3

Source Data Extended Data Fig. 4

Source Data Extended Data Fig. 5

Source Data Extended Data Fig. 6

Source Data Extended Data Fig. 8

Source Data Extended Data Fig. 9

Source Data Extended Data Fig. 10

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing,

adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

# About this article

## Cite this article

Solier, S., Müller, S., Cañeque, T. *et al.* A druggable copper-signalling pathway that drives inflammation. *Nature* **617**, 386–394 (2023). https://doi.org/10.1038/s41586-023-06017-4

- Received: 21 March 2022
- Accepted: 27 March 2023
- Published: 26 April 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-06017-4

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

# This article is cited by

- **[Copper clampdown alleviates inflammation]**

    - Katie Kingwell

  *Nature Reviews Drug Discovery* (2023)

### Copper catalyses changes in cell state

Nature Research Briefing 26 Apr 2023

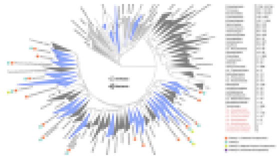
### Copper boosts pro-inflammatory state of macrophages

- Alexandra Flemming

Nature Reviews Immunology Research Highlight 12 May 2023

---

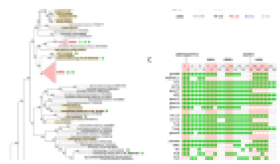
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-06017-4



## Large-scale computational discovery and analysis of virus-derived microbial nanocompartments

06 August 2021

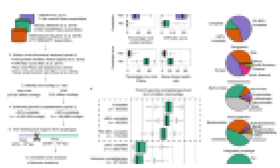
Michael P. Andreas & Tobias W. Giessen



## Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features

16 February 2021

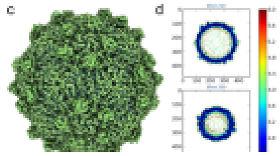
Natalya Yutin, Sean Benler, … Eugene V. Koonin



## Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome

24 June 2021

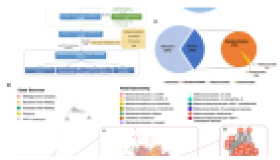
Stephen Nayfach, David Páez-Espino, … Nikos C. Kyrpides



## [Structure of Nora virus at 2.7 Å resolution and implications for receptor binding, capsid stability and taxonomy]

12 November 2020

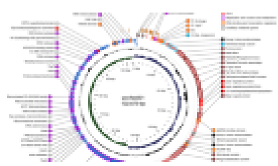
Pasi Laurinmäki, Shabih Shakeel, … Sarah J. Butcher



## [Metagenomic analysis reveals unexplored diversity of archaeal virome in the human gut]

29 December 2022

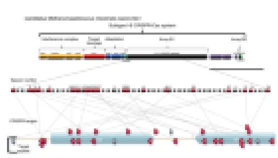
Ran Li, Yongming Wang, … Yingfei Ma



## [ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis]

14 November 2018

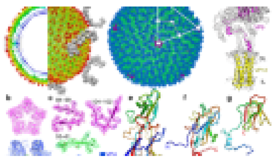
Andrey N. Shkoporov, Ekaterina V. Khokhlova, … Colin Hill



### CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans

17 January 2019

César Díez-Villaseñor & Francisco Rodriguez-Valera



### Assembly of complex viruses exemplified by a halophilic euryarchaeal virus

29 March 2019

Luigi De Colibus, Elina Roine, … David I. Stuart



### Cryo-EM structure of ssDNA bacteriophage ΦCjT23 provides insight into early virus evolution

03 December 2022

Nejc Kejzar, Elina Laanto, … Juha T. Huiskonen

- Article
- Open Access
- Published: 03 May 2023

# Structural atlas of a human gut crassvirus


- Oliver W. Bayfield ,
- Andrey N. Shkoporov,
- Natalya Yutin,
- Ekaterina V. Khokhlova,
- Jake L. R. Smith,
- Dorothy E. D. P. Hawkins,
- Eugene V. Koonin,
- Colin Hill &
- …
- Alfred A. Antson

Show authors



*Nature* **volume 617**, pages 409–416 (2023)



- 9170 Accesses
- 154 Altmetric
- Metrics details


## Abstract


CrAssphage and related viruses of the order *Crassvirales* (hereafter referred to as crassviruses) were originally discovered by cross-assembly of metagenomic sequences. They are the most abundant viruses in the human gut, are found in the majority of individual gut viromes, and account for up to 95% of the viral sequences in some individuals[1,2,3,4]. Crassviruses are likely to have major roles in shaping the composition and functionality of the human microbiome, but the structures and roles of most of the virally encoded proteins are unknown, with only generic predictions resulting from bioinformatic analyses[4,5]. Here we present a cryo-electron microscopy

reconstruction of *Bacteroides intestinalis* virus ΦcrAss001[6], providing the structural basis for the functional assignment of most of its virion proteins. The muzzle protein forms an assembly about 1 MDa in size at the end of the tail and exhibits a previously unknown fold that we designate the 'crass fold', that is likely to serve as a gatekeeper that controls the ejection of cargos. In addition to packing the approximately 103 kb of virus DNA, the ΦcrAss001 virion has extensive storage space for virally encoded cargo proteins in the capsid and, unusually, within the tail. One of the cargo proteins is present in both the capsid and the tail, suggesting a general mechanism for protein ejection, which involves partial unfolding of proteins during their extrusion through the tail. These findings provide a structural basis for understanding the mechanisms of assembly and infection of these highly abundant crassviruses.

# Main

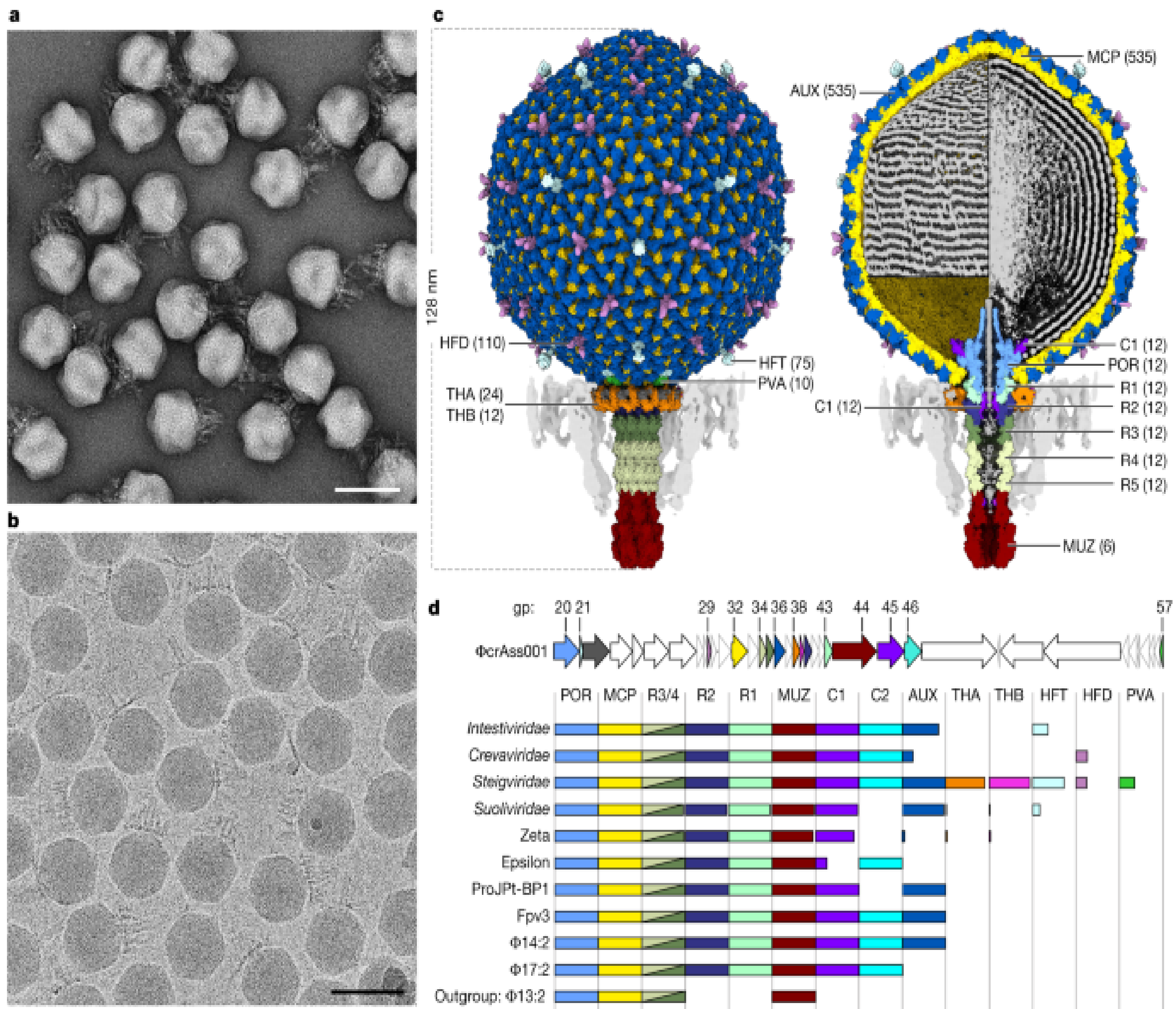
Viruses infecting bacteria are thought to be the most abundant biological entities on earth. Until recently, all such viruses were identified after growing on cultivatable bacterial hosts, but these are likely to account for less than 5% of the actual global diversity[7,8,9]. Recent analyses of viral 'dark matter', made possible by cross-assembly of metagenomic data, have led to the discovery of many new viruses, including crAssphage—the most abundant virus in the human microbiome—and related viruses comprising the newly established order *Crassvirales*[2,4,5,10]. These viruses have been closely associated with human populations throughout evolution[11]. Crassviruses appear to infect exclusively diverse members of the bacterial phylum *Bacteroidota*[3,8,11,12]. In healthy adult Western cohorts, crassviruses are detected in 98–100% of individuals, often dominating the faecal virome[3]. The virus families Flandersviridae, Quimbyviridae and Gubaphage have been identified as close contenders with the *Crassvirales* in terms of their detection frequency[8,12]. However, *Crassvirales* appear to be unique in their showing both a high abundance in individual viromes and across varied cohorts globally[3,8,11,12]. Some crassviruses appear to establish a form of carrier-state infection, with delayed lysis of the infected bacteria, and appear to follow a piggyback-the-winner virus–host dynamic[6,13]. How this lifestyle contributes to their prevalence and abundance remains unclear.

Virus ΦcrAss001 (*Kehishuvirus primarius*) was the first representative of this order to be isolated in pure culture from human faecal samples[6]. Here, we present the characterization of ΦcrAss001 using cryo-electron microscopy (cryo-EM), enabling de novo structure building and elucidation of functions for around 1,440 protein subunits constituting the virion, leading to the identification of a novel 'crass fold' in the distal part of the tail. The structure reveals a mechanism by which a large complement of virally encoded cargo proteins, representing a total molecular mass of approximately 5.5 MDa, can be stored in the capsid and tail. The presence of one such cargo protein inside both the capsid and tail suggests a mechanism for protein ejection that takes place before complete ejection of genomic DNA and involves partial unfolding of cargo proteins. Sequence analysis revealed the extent of conservation of virion proteins throughout the *Crassvirales*, providing a structure–function atlas for this viral order.

## Linking virion and genome architectures

The structure of the ΦcrAss001 virion was determined using single-particle 3D reconstructions from electron microscopy data (Fig. 1a,b). Reconstructions were determined to a resolution that enabled building atomic models for most of the proteins that constitute the virion (Fig. 1c). The virion possesses a short non-contractile tail extending approximately 44 nm from an icosahedral capsid. Inside the capsid, double-stranded DNA (dsDNA) is packed with a left-handed supercoiling (Fig. 1c, left), and at high density leading to hexagonal close-packing (Fig. 1c, right). Moreover, the reconstruction reveals the structural hallmark proteins common to all crassviruses, and sequence analysis enabled mapping of the virion proteins that are most broadly conserved across this viral order (Fig. 1d). In particular, no homologues of the capsid auxiliary protein and the tail muzzle protein were detected among the currently characterized structures of viral proteins.

**Fig. 1: Structure of the ΦcrAss001 virion and functional assignments.**



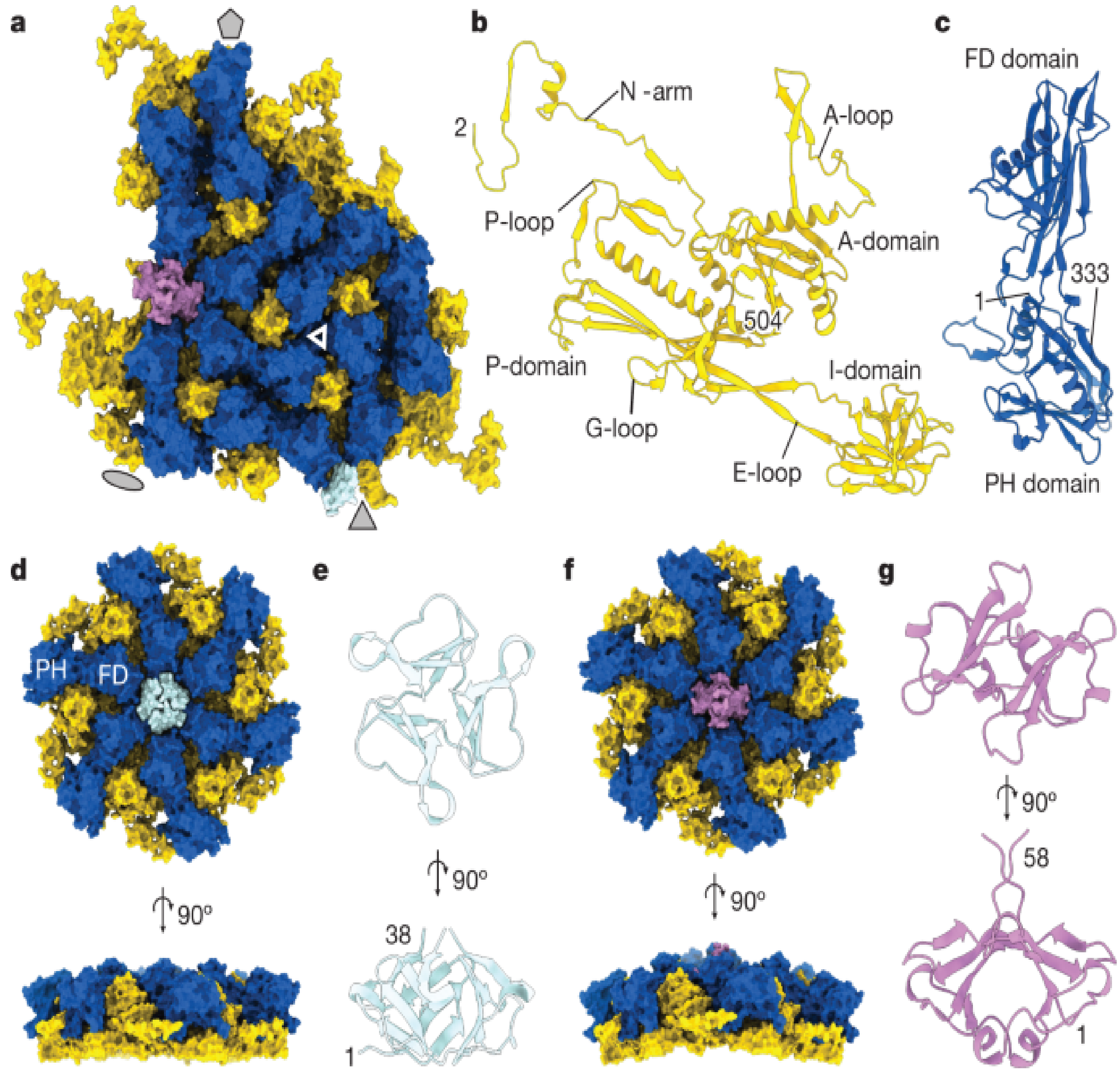

**a**,**b**, Electron micrographs of virions recorded using negative staining (**a**; *n* = 16 micrographs) and under cryogenic conditions (**b**; *n* = 44,006 micrographs). Scale bars, 100 nm. **c**, Molecular surface of virion reconstruction viewed from outside (left) and *z*-clipped (right), coloured by gene product. DNA (grey) on the *z*-clipped image is depicted either as an outer layer lining the capsid wall (top left) or clipped (right), or is removed (bottom left). Unmodelled regions are displayed as map, lowpass filtered to 8 Å (parts of head fibre proteins and DNA) and to 16 Å (tail fibres, semi-transparent grey). AUX, auxiliary capsid protein; C1 and C2, cargo proteins; MUZ, muzzle protein; POR, portal protein; PVA, portal vertex auxiliary capsid protein; R1–R4, ring proteins; THA and THB, tail hub proteins. Numbers in brackets indicate the protein copy number in the virion. **d**, Top, the ΦcrAss001 genome (accession: NC_049977.1) region 11643–68717 showing open reading frames as arrows coloured corresponding to gene

products (gp) in **c**, scaled according to gene length. Bottom, gene conservation, with the length of each bar representing the fraction of each crassvirus group in the order *Crassvirales* containing a detectable homologue, coloured corresponding to **c**. Names on the left refer to the virus families (*Intestiviridae*, *Crevaviridae*, *Steigviridae* and *Suoliviridae*) and groups (zeta and epsilon) with individual members from non-human gut environments represented by ProJPt-BP1 (termite gut), Fpv3 (fish farm) and Φ14:2, Φ17:2 and Φ13:2 (marine environments).

# The capsid proteins

The icosahedral reconstruction of the capsid, which was determined to a resolution of 3.01 Å (Extended Data Table 1), has an outer diameter of approximately 88 nm. The capsid has a triangulation number (*T*) of nine, with an asymmetric unit comprising nine copies of the major capsid protein (MCP) gp32, nine copies of the capsid auxiliary protein gp36, one copy of the trimeric head fibre protein (HFT) gp21, and two copies of the dimeric head fibre protein (HFD) gp29 (Fig. 2a).

**Fig. 2: ΦcrAss001 capsid proteins.**




**a**, Molecular surface of the asymmetric subunit of the icosahedral capsid with indicated symmetry axes (grey shapes), with major capsid protein gp32 (yellow), capsid auxiliary protein gp36 (dark blue), head fibre proteins gp21 (HFT; pale blue, trimer) and gp29 (HFD; pink, dimer). The triangle indicates a local quasi-$C_3$ symmetry formed by I and PH domains. **b**,**c**, Ribbon diagrams of the major capsid protein subunit (**b**) and the capsid auxiliary protein gp32 (**c**). **d**, Molecular surface of the $C_3$ capsid hexon. **e**, Ribbon diagram of the HFT trimer. **f**, Molecular surface of the skewed hexon. **g**, Ribbon diagram of the HFD dimer. Numbers indicate terminal residues of the model.

The MCP adopts the HK97-like fold—the canonical fold of the capsid proteins of the viruses in the class *Caudovirecetes*[14], with the main body comprising the A-domain and P-domain (Fig. 2b). A β-barrel insertion in the E-loop, composed of residues 79–180 (I-domain), is a widespread feature of the crassviruses[5]. This insertion domain forms a three-fold symmetric assembly with the auxiliary protein at the capsid surface (Fig. 2a and Extended Data Fig. 1), making up much of the exposed surface of the capsid and contributing substantially to the capsid wall thickness and stability. An insertion at a similar position of the E-loop, albeit of a different fold (comprising one helix and six strands), is present in the major capsid protein of bacteriophage T4[15]. The ΦcrAss001 I-domain is structurally similar to the I-domain of the major capsid protein of phage P22 (DALI *Z*-score of 6.1 and root mean square deviation (r.m.s.d.) of 2.7 Å with Protein Data Bank (PDB) entry 5UU5 chain E). However, in P22, this domain (residues 226–344) is part of the A-domain, rather than the E-loop.

The auxiliary protein gp36 has a two-domain architecture (Fig. 2c). The PH domain is composed of residues 1–74 and 238–333 and shows structural similarity to the pleckstrin homology-like domain (DALI *Z*-score of 5.7 and r.m.s.d. of 3.2 Å with PDB entry 4DX9, belonging to CATH superfamily 2.30.29.30). The FD domain is composed of residues 75–237 and is structurally similar to the *Helicobacter pylori* flagellar cap protein FliD (DALI *Z*-score 7.3 and r.m.s.d. 3.4 Å with PDB entry 6IWY chain A) and *Pseudomonas aeruginosa* flagellin FliC (DALI *Z*-score 6.7 and r.m.s.d. 3.2 Å with PDB entry 4NX9 chain A). To our knowledge, auxiliary proteins with such a fold and size have not been observed in viral capsids, and they represent a unique feature of most of the crassviruses (Fig. 1d).

The FD domain of the auxiliary protein interacts with the A-domain of the MCP, around the periphery of each hexamer and pentamer. Each hexamer can adopt one of two conformations: winged or planar (Fig. 2d,e). Winged hexons span the edges of the icosahedron, on both sides of the icosahedral two-fold symmetry axes, whereas planar hexons are positioned at the three-fold symmetry axes. At the centre of each hexon is a pocket formed by the MCP and auxiliary protein, which serves as a binding site for fibre proteins. A trimer of the head fibre protein gp21 (residues 1–38) binds in the planar hexon pocket at the three-fold symmetric site (Fig. 2d), whereas a dimer of

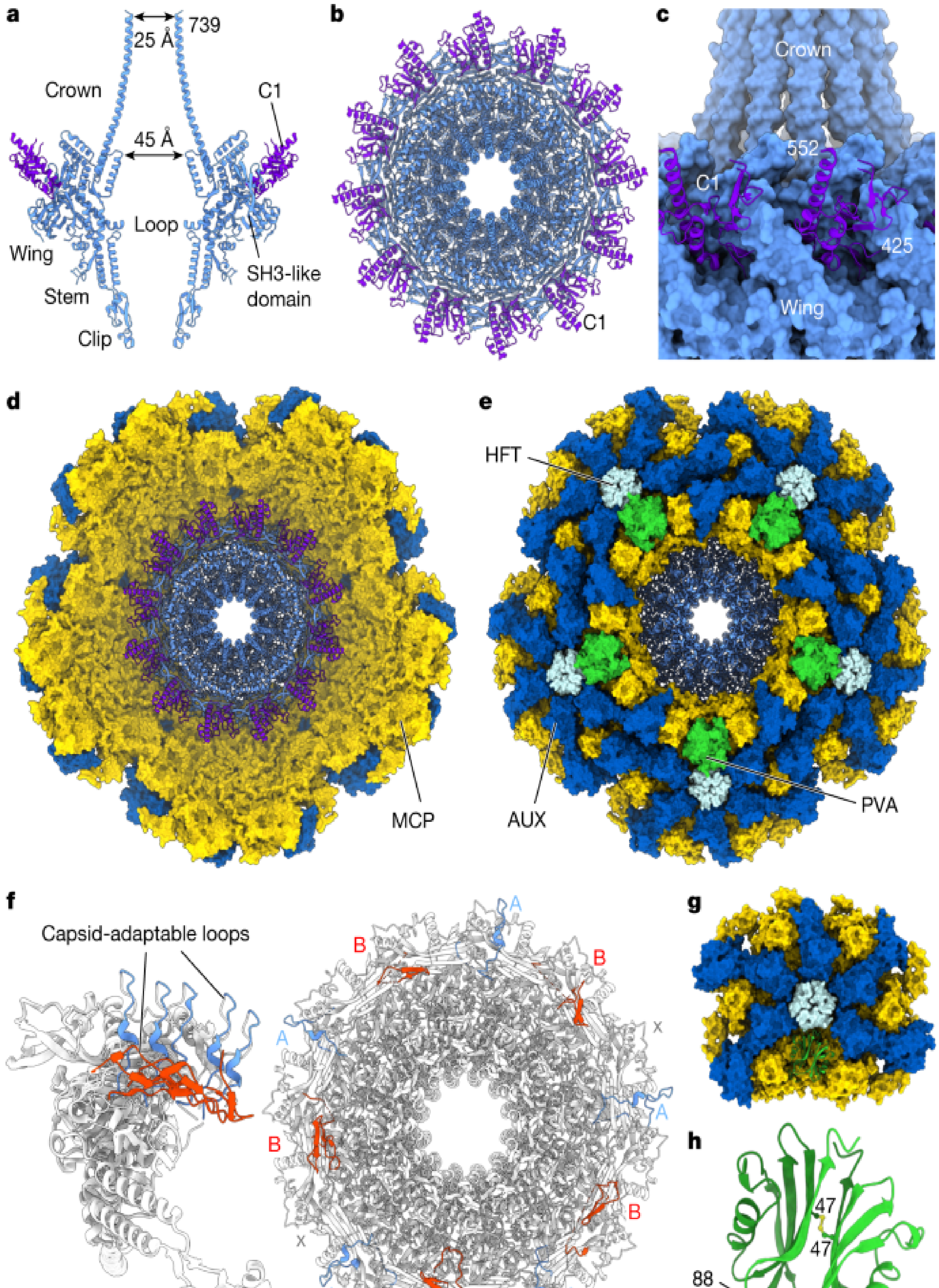
the capsid fibre protein gp29 (residues 1–58) binds in the winged hexon (Fig. 2e). Thus, the *T* number of the capsid is responsible for generating two distinct hexon-pocket environments, which are specific to two different oligomeric fibres. These two fibres are likely to afford two different binding specificities to external molecules located either on the surface of the bacterial host cell or on the surface of the gut. Homologues of these fibres are present in *Intestiviridae* (alpha), *Steigviridae* (beta) and *Suoliviridae* (delta) groups of crassviruses in the case of the HFT gp21, and in the *Steigviridae* (beta) and *Crevaviridae* (gamma) groups in the case of the HFD gp29 (Fig. 1d). Homologues of the C-terminal domain of gp21 (residues 42–97) with high sequence similarity were detected in viruses outside of the *Crassvirales*, including gp8.5 of *Bacillus* phage ϕ29 (40% identity with residues 226–280, and DALI Z-score of 4.3 and r.m.s.d. of 3.2 Å between PDB 3QC7 chain A and gp21 AlphaFold structure prediction) (Extended Data Fig. 2). Gp8.5 has been suggested to bind to cell wall factors of Gram-positive bacteria[16]. AlphaFold modelling showed that the C-terminal domain of gp29 (residues 58–118) forms a β-sandwich (Extended Data Fig. 2) with structural similarity to the immunoglobulin-like fold (DALI *Z*-score of 6.8 and r.m.s.d. of 2.1 Å with PDB 2NSM chain A; CATH superfamily 2.60.40.1120). A DALI search for the N-terminal domains of gp21 and gp29 did not detect any significantly similar structures. Although most groups of crassviruses did not appear to contain homologues of gp21 and gp29, as yet uncharacterized proteins might perform analogous roles.

# The portal vertex

One vertex of the capsid contains the dodecameric oligomer of portal protein gp20. The portal protein is considered the nucleation point for capsid assembly and serves as the docking site for the tail[17,18,19,20]. The portal is surrounded by the five-fold symmetric capsid vertex, contacting 10 major capsid protein subunits. The majority of the portal protein oligomer exhibits 12-fold symmetry and the localized reconstruction imposing $C_{12}$ symmetry enabled structure determination at 3.09 Å resolution (Fig. 3). The majority of the portal protein residues, except for some flexible loop segments (33–35, 68–71, 269–324 and 550–563), could be confidently modelled (Fig. 3a,b). The portal protein exhibits a canonical fold[21] (Fig. 3a), but with a long C-

terminal barrel of the oligomer resembling the corresponding domain of the P22 portal protein[22].

**Fig. 3: The portal protein and the surrounding capsid vertex.**


a
25 Å
739
Crown
C1
45 Å
Loop
Wing
SH3-like domain
Stem
Clip
b
C1
c
Crown
552
C1
425
Wing
d
MCP
e
HFT
PVA
AUX
f
Capsid-adaptable loops
A
B
x
g
h
47
47
88
25

**a**, Two opposing subunits of the portal protein (light blue) and the C1 cargo protein (purple) depicted as ribbons. **b**, Dodecameric assembly of the proteins shown in **a**, viewed along the channel axis. **c**, The portal protein wing region (light blue) interacting with C1 cargo protein subunits (purple). **d**,**e**, Molecular surface of the portal-containing capsid vertex, comprising major capsid protein gp32 (yellow), auxiliary capsid protein gp36 (dark blue) and portal protein gp20 (ribbon, light blue), viewed from inside the capsid along the central channel (**d**) and from outside the capsid (**e**) with the portal vertex auxiliary protein gp57 in green and the head fibre protein gp21 in pale blue. **f**, Ribbon diagram of the portal protein, depicting 12 subunits superimposed (left) and the oligomer viewed along the central axis (right) with capsid-adaptable loops shown in light blue and red corresponding to conformations A and B, respectively. **g**, Unit of the $C_5$ reconstructed portal-containing capsid vertex, with the gp57 (PVA) protein dimer shown as green ribbon. **h**, The gp57 dimer with inter-subunit disulfide bonds show in yellow. Numbers indicate terminal residues of the model.

At the top of the portal protein, an extended α-helical barrel (C-terminal residues 672–739) extends around 80 Å towards the core of the capsid (Fig. 3a). This extended feature has an internal van der Waals diameter of around 45 Å at the base of the crown and around 25 Å at its exposed end. This barrel is well positioned to direct DNA into the capsid during packaging, so that DNA enters at the central part of the capsid towards the opposite capsid wall.

The main body of the portal protein beneath the Crown domain is made up of the stem domain, composed of α-helices defining its central tunnel, and the wing domain (Fig. 3a). Tunnel loops, which are not defined in maps owing to their apparent flexibility, connect the inner stem helix (residues 519–541) with a 34-residue-long kinked helix that is part of the wing (Fig. 2b). These tunnel loops (residues 550–563) are likely to form a narrow constriction in the portal tunnel, as observed in portal proteins of other viruses[18,21,23,24,25].

At the surface of the portal wing, there is clear density corresponding to residues 425–552 of the gp45 cargo protein C1 (Fig. 3b,c). Twelve such segments of gp45 interdigitate with wing loops (SH3-like domains) of the portal, residues 337–380 (Fig. 3b,c).

The 12-fold symmetrical portal oligomer sits within the 5-fold symmetrical capsid vertex (Fig. 3d,e). The localized asymmetric reconstruction of the portal–capsid vertex (obtained without imposing symmetry), determined at 3.60 Å resolution, showed that a loop at the peripheral area of the wing domain, residues 268–319, is flexible and can adopt two preferred conformations (types A and B) in interacting with the $C_5$-symmetric capsid vertex (Fig. 3f). In the type A conformation, residues 277–290 of this 'capsid-adaptable loop' are oriented radially away from the portal's tunnel axis, whereas in the type B conformation residues 292–306 of the loop are arranged tangentially to the portal's circumference (Fig. 3f). In 2 of the 12 subunits of the portal protein, these loops were not resolved in either conformation (denoted 'x'), resulting in the A-B-x-A-B-A-B-A-x-B-A-B order of the conformational states, counting clockwise if viewed from outside of the capsid (Fig. 3f). The result is a pseudo-five-fold symmetrical arrangement of capsid-adaptable loops around the 12-subunit oligomer, assisting the formation of optimal interactions with the five-fold symmetric capsid vertex.

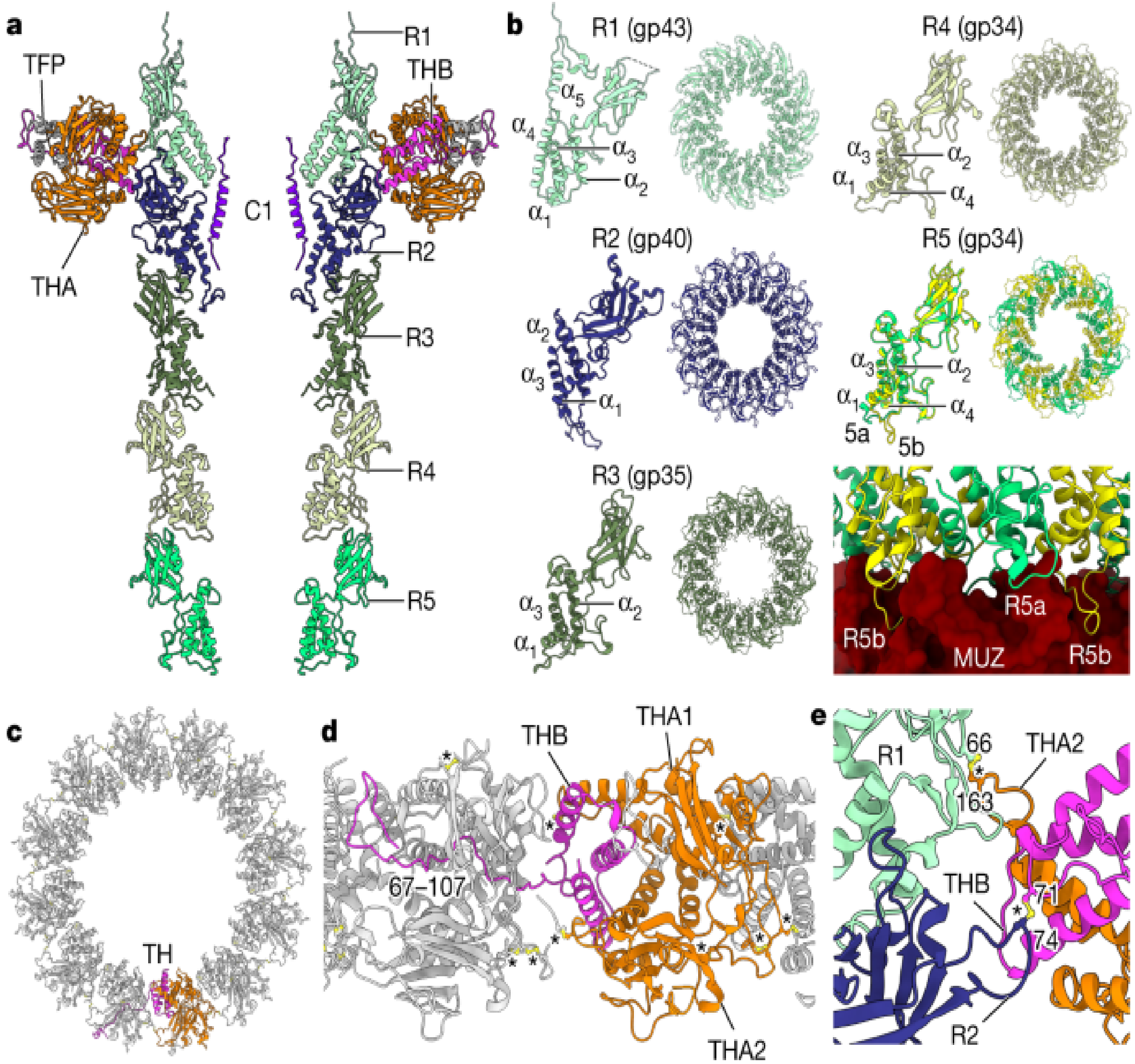
The surrounding capsid vertex was also reconstructed with $C_5$-symmetrical averaging to a resolution of 3.30 Å (Fig. 3g). The portal oligomer positioned in the vertex occludes the binding site of the capsid auxiliary protein gp36, which is substituted by two copies of the portal vertex specific auxiliary protein, gp57 (Fig. 3g,h). The gp57 dimer has approximate two-fold symmetry, reinforced by disulfide bonds between residues 47–47 and 25–88 of the two subunits (Fig. 3h). Notably, hexameric capsomers bordering the unique vertex contain trimer capsid fibres of gp21 (Fig. 3e), in contrast to the hexamers bordering other pentameric vertices, which contain dimeric capsid fibres of gp29.

# Tail barrel assembly and fibre hubs

The portal protein's clip domain interacts with the tail barrel, which extends approximately 23 nm towards the muzzle protein complex at the tip of the tail (Fig. 4). The barrel was reconstructed by imposing $C_{12}$ symmetry to a resolution of 3.20 Å, and with $C_6$ symmetry to a resolution of 3.30 Å. The barrel is composed of five stacked dodecameric 'rings', termed ring 1 (R1, at

the portal end) to ring 5 (R5, at the muzzle end) (Fig. 4a). These structures are formed by four paralogous ring proteins: R1, R2 and R3 are encoded by genes gp43, gp40 and gp35, respectively, whereas R4 and R5 are both encoded by gp34 (Fig. 4b). Such a multi-ring barrel structure results in ΦcrAss001 having a much longer tail than typically observed in other podoviruses, which have only been observed to possess one such ring formed of one protein. The ring proteins share highly similar folds and consist of two domains, with three structurally conserved α-helices and a β-sandwich domain (Fig. 4b). In R5 (gp34), residues 15–26 adopt 2 conformations—5a and 5b—that facilitate binding to the 6-fold symmetric muzzle complex (Fig. 4b). Despite their conserved fold, the sequence identity among the ring proteins is relatively low, ranging from 14.9% (R2–R3) to 18.6% (R1–R2), although the more closely related paralogues in ΦcrAss001—R3 and R4—are 30.6% identical.

**Fig. 4: The tail barrel and fibre docking hubs.**




**a**, Two opposing chains of ring proteins of the tail barrel, depicted as ribbon diagram, from top to bottom: R1 (gp43) (light cyan), R2 (gp40) (navy), R3 (gp35) (mid green), R4 (gp34) and R5 (gp34) (light green), THA (gp38) (orange), THB (gp39) (pink) and gp22 (dark grey). **b**, Individual subunits of each ring protein alongside their dodecameric ring assemblies viewed along the central axis. **c**, The collar formed by 12 tail hub assemblies, viewed along the central axis from the capsid end, depicted in ribbon diagram with chains of one tail hub coloured orange for THA1 and THA2 and pink for THB. TH, tail hub. **d**, The collar as in **c**, rotated 90° and enlarged, showing interlocking tail hubs. **e**, Enlarged view of disulfide bonds (yellow, ball and stick) between tail hub proteins and R1 and R2 proteins. Asterisks designate disulfide bonds.

The α-helical domain of gp34 shares a similar fold with tail-joining factors in tailed bacteriophages with different tail morphologies (HK97, Mu and P22) and in gene-transfer agents[26,27]. The presence of multiple ring protein genes, however, appears to be a conserved and defining feature of crassviruses (Fig. 1d). Some variation exists in this pattern of conservation: genes 34 (R3) and gp35 (R4 and R5) of ΦcrAss001 are more recent paralogues, and many other crassviruses instead encompass a single homologue of R3, R4 or R5, and therefore together with R1 and R2 possess three ring protein genes. The extended tail barrel in ΦcrAss001 that results from this set of ring proteins may assist the tail muzzle in reaching the outer cell membrane, through the approximately-30 nm compact capsular polysaccharide[13]. Having three ring protein genes rather than four could result in a shorter tail barrel, and the outgroup in sequence analysis, phage Φ13:2, is distinctive in its possessing only the R3/R4 gene homologue (Fig. 1d).

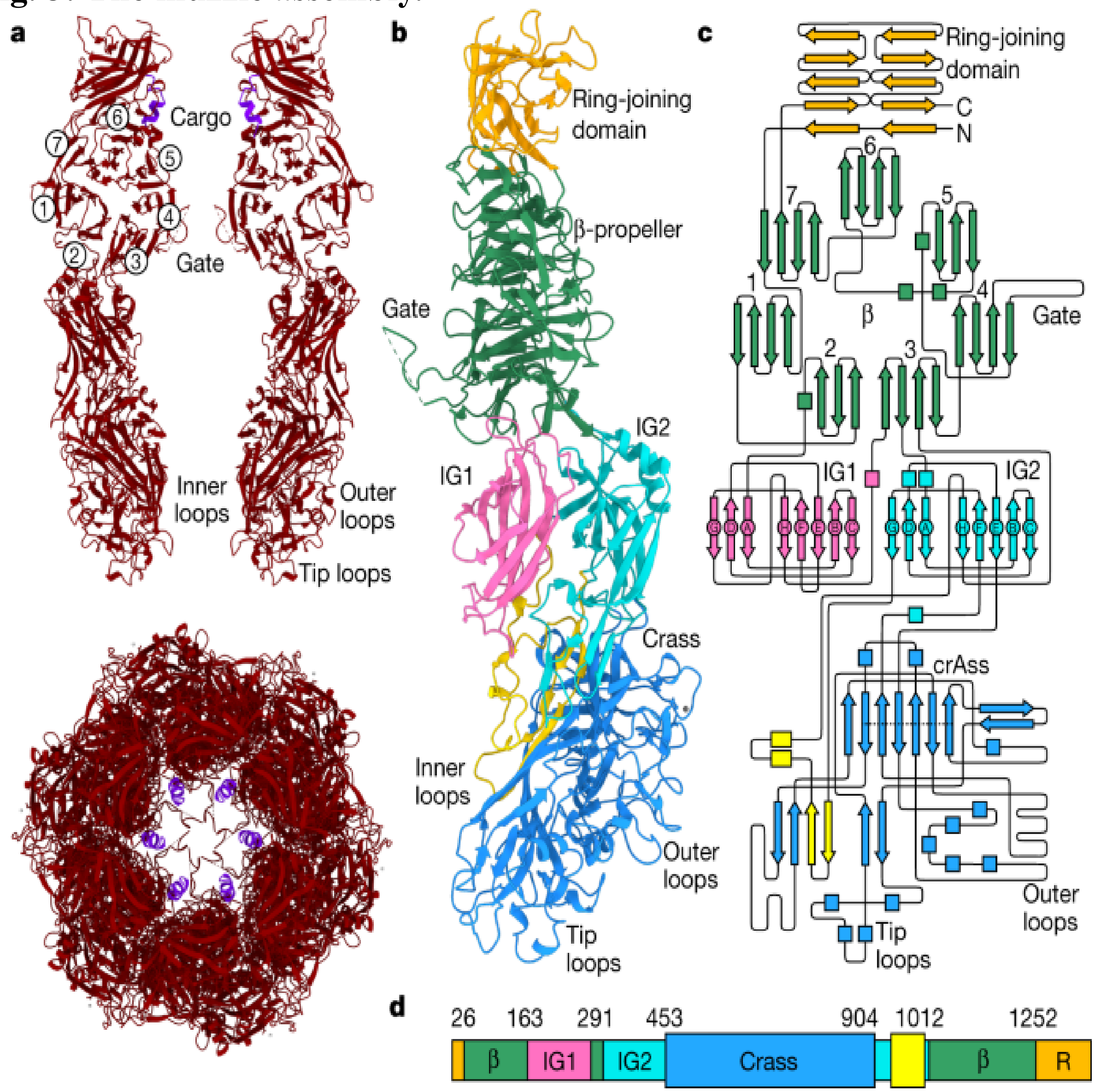
R1 and R2 form an interface for interaction with the tail fibre docking hubs, which are the point of attachment of the tail fibre appendages (Fig. 4a,c). The tail hub (TH) proteins form a continuous ‘collar’ around the tail barrel (Fig. 4c). Each hub assembly is an $A_2B$ trimer that consists of two copies of THA (gp38) and one copy of the smaller protein THB (gp39) (Fig. 4d). The C-terminal segment of THB, residues 67–107, extends into the neighbouring hub, resulting in continuous interlocking ring of 12 docking hubs around the barrel (Fig. 4d). Each docking hub, in turn, interacts with the trimeric tail fibre protein gp22, accommodating its N-terminal residues 1–26 (Fig. 4a). Gp22 interacts with additional tail fibre proteins making up the inner and outer tail fibre assemblies, which are highly flexible and less well resolved in localized reconstructions. The tail barrel and fibre docking hubs are stabilized by disulfide bonds (Extended Data Table 2), notably between fibre docking hub proteins (Fig. 4d), the fibre docking hub and ring proteins (Fig. 4e), and between R1 and R2.

## The muzzle assembly

At the end of the tail is a large hexameric assembly of the muzzle protein gp44 (Fig. 5). The muzzle protein is composed of five domains (Fig. 5a,b), of which the ring-joining domain and the seven-bladed β-propeller are

similar with those of the phage T7 nozzle protein[19], and the remaining three domains are distinct. Below the β-propeller are two immunoglobulin-like domains, and the fifth domain comprising the bulk of gp44 has no counterpart in known virus tail proteins or in any structurally characterized proteins.

**Fig. 5: The muzzle assembly.**




**a**, Top, two opposing subunits of the muzzle protein shown as ribbons (red). Bottom, a hexamer of the muzzle viewed along the central axis with cargo protein in purple. Circled numbers denote the blades of the β-propeller. **b**, A single subunit of the muzzle protein (right chain in **a** rotated by 90° around

the central tunnel axis), with the ring-joining domain in orange, the β-propeller domain in green, the immunoglobulin-like domain 1 (IG1) in pink, IG2 in cyan and the crass domain polypeptide segments in light blue and yellow. **c**, Topology diagram of the muzzle protein with domains coloured as in **b**. The dotted line in the crass domain indicates the β-barrel strands. **d**, Schematic of the muzzle protein with individual domains coloured as in **b**,**c**, with domain boundary residue numbers indicated. β, β-propeller domain. R, ring-joining domain.

Inserted between blades 2 and 3 of the propeller is immunoglobulin-like domain 1 (IG1, residues 163–292) and inserted into blade 3 is an immunoglobulin-like domain 2 (IG2); IG1 and IG2 share a similar topology (Fig. 5b). This immunoglobulin-like fold is structurally similar to the C2-set domains of immunoglobulins[28] (CATH superfamily 2.60.40.150), albeit with a different order of strands, with highest structural similarity observed for synaptotagmin-1[29] (DALI *Z*-score of 5.1 and r.m.s.d. of 3.9 Å with PDB 5CCH chain F).

Serving as a platform for assembly of the large domain forming the distal part of the muzzle is IG2 (between residues 318–1021). This domain is formed by insertions into loops between IG2 strands E–F (residues 453–904) and G–H (residues 940–1012), with the E–F insertion making up the bulk of this region (Fig. 5c). As this domain adopts a fold with no significant structural similarity to known protein domains, we refer to it here as the 'crass fold'. A small β-barrel below domain IG2 (indicated by a dotted line on Fig. 5c; 70 residues spanning segments 453–462, 563–571, 577–580, 581–591, 709–716, 725–733, 833–839, 894–901 and 902–905) structurally resembles other 6-stranded β-barrels (for example, r.m.s.d. of 3.2 Å with 50S ribosomal protein L2, PDB code 4U67, chain B).

The muzzle is gated by a loop within blade 4 of the propeller (Fig. 5a,b). One role of the muzzle is likely to be the formation of this constriction, preventing leakage of DNA and proteins contained in the capsid and tail until infection is initiated. Owing to the restricted channel diameter above and below the gate, and interactions at the inter-subunit interfaces, opening it would require conformational rearrangements of the IG and crass domains. The crass domain, owing to its location on the tail tip and the presence of multiple exposed surface loops, is likely to have a role in contacting the cell

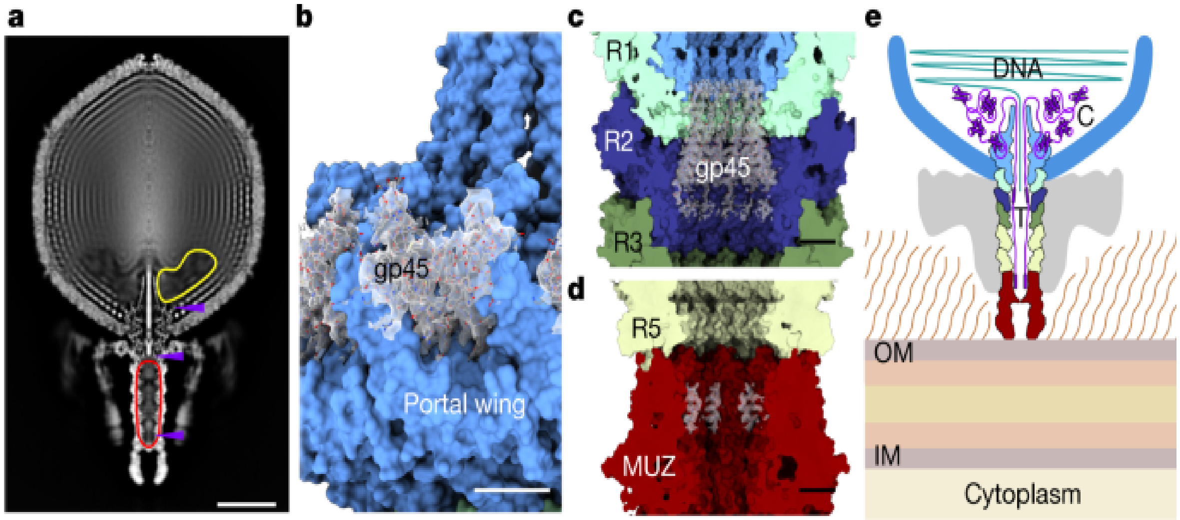
surface, thus initiating these conformational changes. The muzzle protein is highly conserved among the crassviruses (Fig. 1d), but exhibits length variations resulting from insertions into the crass domain and IG2 domain regions that can affect the assembly's surface properties.

## Cargo protein storage zones

Crassviruses are expected to carry several viral gene products inside the virions, including the large multi-subunit RNA polymerase[30], collectively referred to here as cargo proteins. Consistent with this expectation, mass spectrometry of ΦcrAs001 virions (Extended Data Table 3) detected the presence of proteins referred to here as cargo protein 1 (gp45, 90.8 kDa), cargo protein 2 (gp46, 60.2 kDa), and three RNA polymerase-associated proteins: gp47 (262.5 kDa), gp49 (154.5 kDa) and gp50 (276.9 kDa), which are broadly conserved among the crassviruses (Fig. 1d).

To investigate how these cargo proteins pack inside the virion, despite the absence of clear corresponding density in reconstructions, rotational averaging of the virion map around the axis of the portal protein tunnel was performed (Fig. 6a). Inspection of this map revealed an area inside the capsid with a substantially lower density where packing of the surrounding layers of DNA was discontinuous. This internal capsid region located close to the portal vertex was estimated to be about $1.10 \times 10^7$ Å$^3$ in volume. Also notable is the density inside the tail barrel, which is higher than would be expected from solvent alone, and a strong rod-like feature within the portal (Fig. 6a). The space inside the tail barrel measures approximately $8.15 \times 10^5$ Å$^3$.

**Fig. 6: The cargo protein zones.**




**a**, *Z*-clipped view of the rotationally-averaged virion map. The protein cargo zones are indicated by outlines inside the capsid (yellow) and in the tail barrel (red). Purple arrows indicate locations of structured regions of cargo proteins that are resolved in reconstructions. **b**–**d**, Structures of the portal-bound domain of C1 (gp45) (**b**), the tail barrel-bound region of C1 (**c**) and the muzzle-bound cargo protein fragment (**d**) are shown with the corresponding density maps (transparent grey). **e**, Schematic of cargo protein locations inside the virion. Proteins (purple) are stored in the cargo zones located in the tail (T) and capsid (C). IM, inner membrane; OM, outer membrane.

In the cryo-EM reconstructions, fragments of cargo protein 1 (C1) (gp45), were resolved inside both the capsid and the tail. As noted, residues 425–552 fold into a compact domain bound to the portal protein inside the capsid (Fig. 6b). Unexpectedly, well-defined density at the inner surface of R1 and R2 of the tail barrel corresponds to 12 copies of an additional α-helical segment of the same protein, gp45 (residues 214–245) (Fig. 6c). Thus, different portions of gp45 are found inside either the capsid or the tail, separated by the portal protein tunnel.

On the basis of volume measurements, we can estimate the storage capacity of the cargo zones. Assuming proteins were packed at 2.15 $Å^3$ per Da volume per molecular weight, consistent with typical protein crystals with around 43% solvent content[31], the capsid and tail cargo zones could store proteins with a total molecular mass of up to around 5.1 MDa inside the

capsid and an additional 0.4 MDa inside the tail barrel. These two zones can thus harbour multiple copies of the cargo proteins, as detected by mass spectrometry (for example, 6 copies of each the three RNA polymerase-associated protein and up to 12 copies of the smaller proteins gp45 and gp46). The remaining internal capsid volume available for the 102,679 bp genome is approximately $1.84 \times 10^8$ Å$^3$, resulting in a DNA packing density of 0.56 bp nm$^{-3}$, consistent with packaging densities observed for other viruses[32].

A further region of density inside the tail was observed at its distal end, next to the ring-joining domain of the muzzle protein (Fig. 6d). This region corresponds to 6 copies of a 14-residue segment containing an α-helix. Although this segment could not be assigned to a specific protein, it probably corresponds to yet another N-terminal segment of gp45, extending from the 214–245 fragment. Indeed, peptides spanning the regions 79–101 and 181–190 of gp45 were detected by mass spectrometry (Extended Data Fig. 3). Consistent with this, the $8.15 \times 10^5$ Å$^3$ volume of the tail cargo zone could accommodate up to about 0.4 MDa of protein material, which is sufficient for 12 copies of residues 1–245 of gp45 (about 306 kDa). This makes the capsid cargo zone the likely locus for the remaining cargo proteins, C2 (gp46) and the RNA polymerase-associated proteins (gp47, gp49 and gp50) (Fig. 6e).

# A model for protein ejection

The data reported here, along with previous observations on phage P22, and other podoviruses including T7, provide the basis for proposing a hypothetical model for protein ejection (Extended Data Fig. 4). Although podoviruses do not have a tail barrel as extended as that of ΦcrAss001, many do contain ‘ejectosome’ proteins that are packed as cargo inside their capsids. In the bacteriophage P22 capsid, there are 3 such proteins[33] (gp7, gp16 and gp20) with a combined mass of around 2.5 MDa. For T7, ejection proteins were shown to form a membrane-spanning channel during the initial phase of infection[34]. Although not all cargo proteins were resolved in the reconstructions of ΦcrAss001, these proteins have to exit the capsid ahead of DNA because they form the channel for DNA ejection. As in the case of T7, we expect that gp45 and gp46 of ΦcrAss001 are involved in

transmembrane channel formation, with the other three cargo proteins annotated as RNA polymerase-related proteins[5]. Given that one region of gp45 (residues 214–245) is already poised for ejection, positioned inside the tail barrel of ΦcrAss001, this protein is likely to be ejected first and to contribute to channel formation. A transmembrane helix is predicted by TMHMM-2.0[35] to exist in the N-terminal region of gp45 (residues 136–158) (Extended Data Fig. 3), which could insert into the host cell membrane in a similar manner as described for the T7 protein gp14[36].

The two parts of cargo protein C1, gp45, identified in the tail barrel (residues 214–245) and in the capsid (residues 425–552) are separated by a distance of approximately 420 Å and could be bridged by the connecting 179 residues. Indeed, peptides spanning the regions 275–300 and 307–320 of gp45 were detected by mass spectrometry (Extended Data Fig. 3), indicating that these regions are present inside the virion. We can expect gp45 to be ejected in the N→C direction: with the N-terminal part located in the tail exiting first followed by the C-terminal domain located in the capsid (Extended Data Fig. 4). The size and globular structure of the C-terminal part of gp45 (Fig. 6b) suggests that its ejection involves partial unfolding.

The other four proteins detected by mass spectrometry (gp46, gp47, gp49 and gp50) are predicted by AlphaFold2[37] to adopt folded α-helical structures (Extended Data Fig. 5) and consistently, the crystal structure of a homologue of gp49 showed it is predominantly α-helical[30]. Several constrictions inside the tail barrel are less than 50 Å in diameter, requiring that these cargo proteins are partially unfolded to pass through.

Although understanding precisely how proteins and DNA are ejected requires further studies, it is likely that ejection is facilitated by pressure inside the head, which can reach[38] around 100 atmospheres. The order in which crassvirus proteins are extruded through the tail is likely to be controlled both by the way these proteins are packed and by their interactions with one another. Once the cargo proteins forming the transmembrane tunnel have exited the virion, DNA could begin to descend through the tail barrel. DNA ejection could be facilitated by solid-to-fluid–like transitions as suggested for bacteriophage λ[39,40].

When considering the roles of the cargo proteins, it should be noted that members of the epsilon group of *Crassvirales* lack a detectable homologue of gp45, whereas homologues of gp46 are missing in the *Suoliviridae* (delta) and zeta groups (Fig. 1d). Thus, the sets of cargo proteins and accordingly the ejection mechanisms may vary across the crassviruses.

In conclusion, the reconstruction of the ΦcrAss001 virion identifies the structural hallmarks of crassviruses and—combined with comparative genomic analysis—has enabled us to assign functions to the majority of the previously uncharacterized proteins involved in virion assembly and infection. Among these structural features, the muzzle protein containing the novel crass fold functions as a gatekeeper, which probably contributes to the contact between the virion and the host cell surface. The extended tail barrel formed by widely conserved paralogous ring proteins has a high capacity for storage of cargo proteins alongside the capsid cargo zone, ensuring that the proteins required for the initial stages of crassvirus infection are carried inside virions and delivered to the host cell. The presence of a cargo protein storage zone in the tail is a distinct feature that, to our knowledge, has not been observed in viruses before, and suggests a mechanism of compartmentalization of the virus genome and cargo proteins ensuring their ordered release into the host.

# Methods

## Virus propagation and purification

The host strain *B. intestinalis* APC919/174 was propagated anaerobically in Fastidious Anaerobe Broth (Neogen) at 37 °C. Cultures were infected once a value of $A_{600} = 0.2$ (~$2 \times 10^8$ colony-forming units per ml) was reached, at a multiplicity of infection (MOI) of 1, and grown overnight at 37 °C in anaerobic jars. Cultures were centrifuged for 15 min at 5,200*g* at 4 °C and filtered through 0.45-µm polyethersulfone (PES) filters. Filtered lysate was concentrated by polyethylene glycol precipitation by addition of solid PEG-8000 to 10% w/v and NaCl to 0.5 M and incubation overnight at 4 °C. The solution was centrifuged at 5,000*g* for 30 min at 4 °C, and the pellet resuspended in a minimum volume of SM buffer (100 mM NaCl, 10 mM

$MgSO_4$, 50 mM Tris-HCl pH 7.5). The solution was applied to the top of a caesium chloride step gradient in an Ultra-Clear tube (Beckman), comprising of solutions of densities ~1.31 and ~1.55 g $ml^{-1}$ in 20 mM Tris-HCl pH 7.5, 10 mM $MgSO_4$. Gradients were centrifuged at 28,000 rpm in an SW28 rotor at 4 °C for 3 h. The band containing mature virions was extracted by needle side-puncture. Samples were dialysed against buffer containing 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 10 mM $MgSO_4$ overnight at 4 °C. Virus particles were further concentrated by centrifugation at 40,000 rpm in an SW41 Ti rotor at 8 °C for 1 h and resuspending in 50 μl of buffer. This sample was used for all subsequent electron microscopy analysis.

## Electron microscopy

For negative staining, carbon-formvar coated copper grids (Agar Scientific) were plasma cleaned in a PELCO easiGlow for 60 s at 0.38 mbar (air) and 20 mA. Sample solution (5 μl) was applied to a grid for 1 min then blotted with filter paper, and stained with 5 μl of 2% w/v uranyl acetate. Micrographs were recorded using a Tecnai 12 G2 BioTWIN microscope with tungsten filament operating at 120 kV, with a Ceta 16M camera (Thermo Fisher Scientific). For initial screening of samples, 16 negatively stained micrographs were recorded at magnifications of 9,300–98,000×. For cryo-EM, sample was applied to lacey carbon grids supporting an ultrathin carbon layer on 400-mesh copper (Ted Pella), glow discharged as above, and vitrified using a Vitrobot Mark IV (Thermo Fisher Scientific) at 100% humidity and 4 °C. Data were collected on a 300 kV Titan Krios (Thermo Fisher Scientific) with images recorded on a Falcon3EC detector operating in integrating mode, using EPU software (2.5.0). In total, four datasets were recorded (Extended Data Table 1) and processed together in RELION 3.1[41]. Motion correction was performed using RELION and contrast transfer functions (CTFs) were estimated using CTFFIND4[42] using dose-weighting. Empty micrographs containing no virus particles were excluded from processing beyond CTF estimation. Particles selected after 2D classification were used for 3D refinements, initially using icosahedral symmetrical averaging. To check if magnified pixel size calibrations were consistent across datasets, reconstructions were compared using

determine_relative_pixel_size.py (890eb35)[43] and by manually inspection, and pixel sizes scaled to reference optics group 1. A symmetry-expanded set of particles was generated using relion_particle_symmetry_expand. The unique vertex of the capsid was identified by application of a cylindrical mask on one vertex and 3D classification into 10 classes without performing image alignment. Particles with tails inside the masked region after classification were selected for further rounds of 3D refinement using local searches, after extracting sub-particles using relion_localized_reconstruction[44]. Alternatively, in the case of the muzzle assembly, sub-particles were extracted prior to masked classification. Additional symmetry expansions were performed to resolve the muzzle assembly from tail barrel particles refined with $C_{12}$ symmetry, and of the asymmetric portal-containing vertex. Refinements were followed by CTF refinement of per-particle defocus values and astigmatism, and magnification anisotropy estimation. For the symmetry-free reconstruction of the whole virion, re-extracted particles were enlarged to include the whole of the virus particle images centred on the virion and reconstructed with Ewald sphere correction. Resolutions of the reconstructions were estimated using the Fourier shell correlation 0.143 threshold and using RELION's local resolution estimation. Owing to the large particle size, gold diffraction data collected for optics group 1 was used to check pixel size calibration and magnification anisotropy[45], and the final pixel size scaled during postprocessing. Atomic models (Extended Data Fig. 6) were generated by manual building in Coot (0.9.8.1)[46], followed by real space refinement in Phenix (1.19)[47]. Images were generated using ChimeraX (1.5)[48].

## Mass spectrometry analysis

Two samples of purified virions were analysed: one sample, R, was reduced and the other, NR, was untreated before running on 4–12% SDS–PAGE (Invitrogen), and the Coomassie-stained band was excised. Sample R was further reduced with 1.5 mg ml$^{-1}$ dithioerythritol in-gel, while NR was untreated. Both samples were alkylated with 9.5 mg ml$^{-1}$ iodoacetamide and digested overnight with trypsin. Resulting peptides were analysed by liquid chromatography–tandem mass spectrometry (LC–MS/MS) over 1 h acquisitions with elution from a 50 cm EasyNano PepMap column into an Orbitrap Fusion Tribrid mass spectrometer operated in DDA mode.

Duplicate acquisitions were performed for each sample, one with MS2 measured in the liner ion trap and the other with high resolution MS2 acquired in the Orbitrap mass analyser. Spectra were peak picked and searched against a sequence database containing ΦcrAss001 hypothetical proteins and common proteomic contaminants. Oxidation of methionine and carbamidomethylation of cystine were set as variable modifications. Identified peptides were filtered to 1% false discovery rate as assessed against a reversed database search. Relative quantification between samples and acquisitions was performed within PEAKS Studio (10.5)[49] from areas of aligned extracted ion chromatograms. The mean of the peak areas across the ion trap and Orbitrap acquisitions were calculated for each protein in R and NR samples respectively.

## AlphaFold structure predictions

Structure predictions using AlphaFold2[37] were performed for ΦcrAss001 protein sequences: gp46, gp47, gp49, gp50, gp21 and gp29. In the case of gp21 and gp29, the C-terminal domains that could not be modelled ab initio using electron microscopy reconstructions (owing to limited resolution and flexibility in these parts) were rigid-body docked into the density maps above the corresponding gp21 and gp29 capsid binding sites. Complete pseudo-atomic models for each fibre protein oligomer were created by fusing the N-terminal (cryo-EM modelled) and C-terminal (AlphaFold modelled) regions linked in coot.

## Protein sequence analysis

To calculate the representation of ΦcrAss001 virion protein homologues across the *Crassvirales*, the protein sequences encoded in the previously described reference genome set[5] were searched. The reference set consisted of 673 virus genomes grouped into 211 clusters at 90% similarity at the genome nucleotide sequence level. Each virion protein was queried against the reference set using PSI-BLAST[50]. To account for over-representation of some crassviruses, a genome cluster was given a score of 1 if a ΦcrAss001 protein homologue was found in more than half of the genomes in that cluster (otherwise 0). The cluster scores were then summed within the previously delineated groups of crassviruses (alpha to zeta) and plotted

against the corresponding ΦcrAss001 proteins (Extended Data Table 4, Fig. 1d). To identify homologues outside the *Crassvirales*, the Genbank and RefSeq virus databases were queried using PSI-BLAST.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Cryo-EM reconstructions and atomic coordinates comprising the virion capsid and tail structures have been deposited to the Electron Microscopy Data Bank under accession codes EMD-14088, EMD-14089, EMD-14090, EMD-14091, EMD-14092, EMD-14093, EMD-14094 and EMD-14100 (maps) and to the Protein Data Bank under accession codes 7QOF, 7QOG, 7QOH, 7QOI, 7QOJ, 7QOK and 7QOL (atomic coordinates), with example regions of maps and corresponding models in Extended Data Fig. 6. Data collection and refinement statistics are presented in Extended Data Table 1. Multiple sequence alignments of ΦcrAss001 virion protein homologues are available at https://ftp.ncbi.nih.gov/pub/yutinn/crAss_structural_prot_2022/.

## References


1. Shkoporov, A. N. & Hill, C. Bacteriophages of the human gut: the “known unknown” of the microbiome. *Cell Host Microbe* **25**, 195–209 (2019).

2. Dutilh, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 4498 (2014).

3. Guerin, E. et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* **24**, 653–664.e6 (2018).

4. Yutin, N. et al. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2018).

5. Yutin, N. et al. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat. Commun.* **12**, 1044 (2021).

6. Shkoporov, A. N. et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. *Nat. Commun.* **9**, 4781 (2018).

7. Roux, S. et al. Minimum information about an uncultivated virus genome (MIUVIG). *Nat. Biotechnol.* **37**, 29–37 (2019).

8. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109.e9 (2021).

9. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).

10. Koonin, E. V. & Yutin, N. The crAss-like phage group: how metagenomics reshaped the human virome. *Trends Microbiol.* **28**, 349–359 (2020).

11. Edwards, R. A. et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol.* **4**, 1727–1736 (2019).

12. Benler, S. et al. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome* **9**, 78 (2021).

13. Shkoporov, A. N. et al. Long-term persistence of crAss-like phage crAss001 is associated with phase variation in *Bacteroides intestinalis*. *BMC Biol.* **19**, 1–16 (2021).

14. Wikoff, W. R. et al. Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289**, 2129–2133 (2000).

15. Fokine, A. et al. Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc. Natl Acad. Sci. USA* **102**, 7163–8 (2005).

16. Xiang, Y. & Rossmann, M. G. Structure of bacteriophage ϕ29 head fibers has a supercoiled triple repeating helix–turn–helix motif. *Proc. Natl Acad. Sci. USA* **108**, 4806–4810 (2011).

17. Motwani, T. & Teschke, C. M. Architect of virus assembly: the portal protein nucleates procapsid assembly in bacteriophage P22. *J. Virol.* **93**, 00187-19 (2019).

18. Sun, L. et al. Cryo-EM structure of the bacteriophage T4 portal protein assembly at near-atomic resolution. *Nat. Commun.* **6**, 7548 (2015).

19. Cuervo, A. et al. Structures of T7 bacteriophage portal and tail suggest a viral DNA retention and ejection mechanism. *Nat. Commun.* **10**, 3746 (2019).

20. Chaban, Y. et al. Structural rearrangements in the phage head-to-tail interface during assembly and infection. *Proc. Natl Acad. Sci. USA* **112**, 7009–7014 (2015).

21. Lebedev, A. A. et al. Structural framework for DNA translocation via the viral portal protein. *EMBO J.* **26**, 1984–1994 (2007).

22. Tang, J. et al. Peering down the barrel of a bacteriophage portal—the genome packaging and release valve in P22. *Structure* **19**, 496–502 (2011).

23. Lokareddy, R. K. et al. Portal protein functions akin to a DNA-sensor that couples genome-packaging to icosahedral capsid maturation. *Nat. Commun.* **8**, 14310 (2017).

24. Liu, Y.-T., Jih, J., Dai, X., Bi, G.-Q. & Zhou, Z. H. Cryo-EM structures of herpes simplex virus type 1 portal vertex and packaged genome.

*Nature* **570**, 257–261 (2019).

25. Simpson, A. A., Petr, G., Jardine, P. J., Dwight, L. & Michael, G. Structure determination of the head–tail connector of bacteriophage phi29. *Acta Crystallogr. D* **D57**, 1260–1269 (2001).

26. Iwasaki, T. et al. Three-dimensional structures of bacteriophage neck subunits are shared in Podoviridae, Siphoviridae and Myoviridae. *Genes Cells* **23**, 528–536 (2018).

27. Bárdy, P. et al. Structure and mechanism of DNA delivery of a gene transfer agent. *Nat. Commun.* **11**, 3034 (2020).

28. Leonard, T. A. in *Encyclopedia of Metalloproteins* (eds Kretsinger, R. H., Uversky, V. N. & Permyakov, E. A.) 309–318 (Springer, 2013).

29. Holm, L. DALI and the persistence of protein shape. *Protein Sci.* **29**, 128–140 (2020).

30. Drobysheva, A. V. et al. Structure and function of virion RNA polymerase of a crAss-like phage. *Nature* **589**, 306–309 (2021).

31. Matthews, B. W. Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497 (1968).

32. Hua, J., Huet, A., Lopez, C. A. & Toropova, K. Capsids and genomes of jumbo-sized bacteriophages reveal the evolutionary reach of the HK97 fold. *mBiol.* **8**, 01579-17 (2017).

33. Wu, W. et al. Localization of the houdinisome (ejection proteins) inside the bacteriophage P22 virion by bubblegram imaging. *mBio* **7**, 01152-16 (2016).

34. Swanson, N. A. et al. Cryo-EM structure of the periplasmic tunnel of T7 DNA-ejectosome at 2.7 Å resolution. *Mol. Cell* **81**, 3145–3159.e7 (2021).

35. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov

model: Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

36. Chen, W. et al. Structural changes in bacteriophage T7 upon receptor-induced genome ejection. *Proc. Natl Acad. Sci. USA* **118**, e2102003118 (2021).

37. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

38. Smith, D. E. et al. The bacteriophage φ29 portal motor can package DNA against a large internal force. *Nature* **413**, 748–752 (2001).

39. Evilevitch, A. The mobility of packaged phage genome controls ejection dynamics. *eLife* **7**, e37345 (2018).

40. Liu, T. et al. Solid-to-fluid-like DNA transition in viruses facilitates infection. *Proc. Natl Acad. Sci. USA* **111**, 14675–14680 (2014).

41. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).

42. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).

43. Wilkinson, M. E., Kumar, A. & Casan al, A. Methods for merging data sets in electron cryo-microscopy. *Acta Crystallogr. D* **75**, 782–791 (2019).

44. Ilca, S. L. et al. Localized reconstruction of subunits from electron cryomicroscopy images of macromolecular complexes. *Nat. Commun.* **6**, 8843 (2015).

45. Grant, T. & Grigorieff, N. Automatic estimation and correction of anisotropic magnification distortion in electron microscopes. *J. Struct. Biol.* **192**, 204–208 (2015).

46. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).

47. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).

48. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

49. Tran, N. H. et al. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* **16**, 63–66 (2019).

50. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

# Acknowledgements

The authors thank S. Hart and J. Turkenburg for assistance with cryo-EM; A. Dowle and C. Taylor for assistance with mass spectrometry experiments; R. Thompson, E. Hesketh and D. Maskell; and support staff at eBIC (Diamond Light Source, UK). This project used the Viking Cluster high performance compute facility provided by the University of York; we are grateful for computational support from the University of York High Performance Computing service, Viking and the Research Computing team. This work was supported by the Wellcome Trust (grant 206377, 224665 to A.A.A.), Instruct-ERIC (PID: 10175), and Diamond Light Source for access to eBIC under proposal EM19832 funded by the Wellcome Trust, MRC and BBRSC. The cryo-EM facility at the University of York was supported by the Wellcome Trust, grant 206161. A.N.S. was supported by the Wellcome Trust (grant 220646) and ERC (grant agreement no. 101001684). E. V. Khokhlova and C.H. were supported by Science Foundation Ireland (SFI) under Grant no. SFI/12/RC/2273. N.Y. and E. V. Koonin are supported by the Intramural Research Program of the National Institutes of Health of the USA (National Library of Medicine). Molecular graphics and analyses were

performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

# Author information

## Authors and Affiliations

1. York Structural Biology Laboratory, Department of Chemistry, University of York, York, UK

   Oliver W. Bayfield, Jake L. R. Smith, Dorothy E. D. P. Hawkins & Alfred A. Antson

2. APC Microbiome Ireland and School of Microbiology, University College Cork, Cork, Ireland

   Andrey N. Shkoporov, Ekaterina V. Khokhlova & Colin Hill

3. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

   Natalya Yutin & Eugene V. Koonin

## Contributions

O.W.B., A.N.S., E. V. Koonin, C.H. and A.A.A. conceived the study. A.N.S. and E. V. Khokhlova cultured cells and viruses. O.W.B., A.N.S. and E. V. Khokhlova purified virions. O.W.B. and D.E.D.P.H. collected cryo-EM data. O.W.B. processed cryo-EM data. O.W.B. and J.L.R.S. built and refined atomic models. O.W.B. and A.A.A. analysed structural data. N.Y. and E. V. Koonin performed comparative genomic analysis. O.W.B., A.N.S., E. V. Koonin, C.H. and A.A.A. wrote the manuscript, which was read, edited and approved by all authors.

### Corresponding authors

Correspondence to Oliver W. Bayfield or Alfred A. Antson.

## Ethics declarations

### Competing interests

The authors declare no competing interests.

## Peer review

### Peer review information

*Nature* thanks Jeremy Barr, Antoni Luque and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

## Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Extended data figures and tables

### Extended Data Fig. 1 Capsid wall viewed along the local three-fold symmetry axis.

Molecular surface (left) and diagram of protein layers (right). **(a)** Inner capsid layer formed by three major capsid protein subunits contributed from three adjacent capsomers, orange; **(b)** Middle layer containing three additional overlaying major capsid protein subunits, yellow; **(c)** Outer layer containing three overlaying auxiliary capsid proteins, blue. P, peripheral domain of the major capsid protein. I, insertion domain of the major capsid

protein. PH, pleckstrin-homology domain of the auxiliary capsid protein. G, G-loop of the major capsid protein.

## Extended Data Fig. 2 Models of the head fiber proteins.

**(a–c)**, head fiber gp21 trimer. **(d–f)**, head fiber gp29 dimer. **(a)** Extracted density fitted with the gp21model; shown in two orthogonal views. **(b)** Ribbon diagram with domain structure deduced from cryo-EM density in blue and domain structure predicted by AlphaFold in grey. **(c)** same as B but with only one monomer shown. **(d–f)** same as (a-c) but for gp29, with the domain structure derived from cryo-EM density shown in pink. Scale bars 20 Å.

## Extended Data Fig. 3 Sequence of the cargo protein gp45.

Regions detected by mass spectrometry are in purple, structured regions identified in cryo-EM density are bold and underlined; the predicted transmembrane helix is in yellow.

## Extended Data Fig. 4 Model for ejection of cargo proteins.

Left cargo proteins (purple) are stored in the tail barrel and capsid cargo zones in association with DNA (green). Middle: cargo proteins are being extruded through the constrictions of the portal and tail barrel and are passing through the transmembrane channel (purple) into the host's cytoplasm. Injected proteins can undergo refolding in the cytoplasm. Right: cargo proteins' ejection is complete and is followed by DNA ejection. Viral RNA polymerases have re-folded and transcription of viral genome begins. CPS, capsular polysaccharide.

## Extended Data Fig. 5 Ribbon diagrams of cargo protein structures predicted by AlphaFold.

Top row, coloured according to the pLDDT score: 50, orange; 70, yellow; 100, blue. Mean overall pLDDT is indicated underneath each protein structure. Bottom row, coloured according to secondary structure (α-helix,

red; β-strand, blue; coil, grey) with overall percentages of α- and β- residues indicated underneath.

## Extended Data Fig. 6 Representative density maps with corresponding fitted models for regions of virion proteins.

Maps are depicted in blue mesh with atomic models shown in stick style coloured by atom type. Each protein is labelled with residue ranges shown and the corresponding PDB codes indicated in parentheses.

**Extended Data Table 1 Data collection and refinement**
**Extended Data Table 2 Disulfide bonds stabilising the tail assembly**
**Extended Data Table 3 Virion proteins detected by mass spectrometry**
**Extended Data Table 4 Representation of ΦcrAss001 protein homologues across crassviruses**

# Supplementary information

## Reporting Summary

## Peer Review File

# Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

[Reprints and Permissions](#)

# About this article

## Cite this article

Bayfield, O.W., Shkoporov, A.N., Yutin, N. *et al.* Structural atlas of a human gut crassvirus. *Nature* **617**, 409–416 (2023). https://doi.org/10.1038/s41586-023-06019-2

- Received: 26 July 2022
- Accepted: 27 March 2023
- Published: 03 May 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-06019-2

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

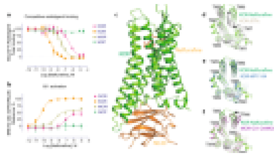
Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

## Structure of the most abundant human gut virus

Nature Research Briefing 03 May 2023

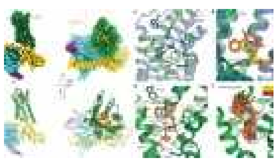
This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-06019-2



## Molecular mechanism of biased signaling at the kappa opioid receptor

11 March 2023

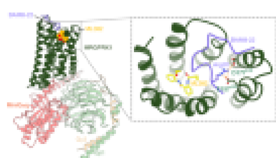
Amal El Daibani, Joseph M. Paggi, … Tao Che



## Structure of the μ-opioid receptor–Gi protein complex

13 June 2018

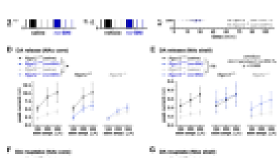
Antoine Koehl, Hongli Hu, … Brian K. Kobilka



## Ligand recognition and allosteric modulation of the human MRGPRX1 receptor

27 October 2022

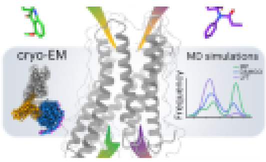
Yongfeng Liu, Can Cao, … Bryan L. Roth



[Role of RGS12 in the differential regulation of kappa opioid receptor-dependent signaling and behavior]

29 May 2019

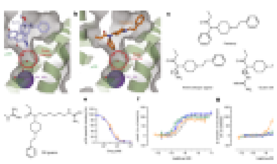
Joshua D. Gross, Shane W. Kaski, … Vincent Setola




[Insights into distinct signaling profiles of the μOR activated by diverse agonists]

21 November 2022

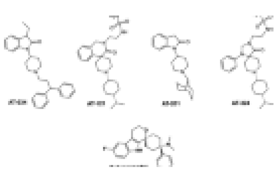
Qianhui Qu, Weijiao Huang, … Georgios Skiniotis



[Structure-based design of bitopic ligands for the μ-opioid receptor]

30 November 2022

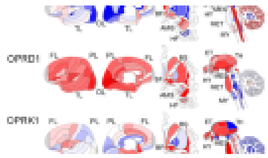
Abdelfattah Faouzi, Haoqing Wang, … Susruta Majumdar



[Attenuated G protein signaling and minimal receptor phosphorylation as a biochemical signature of low side-effect opioid analgesics]

03 May 2022

Pooja Dasgupta, Anika Mann, … Stefan Schulz



[Untangling the complexity of opioid receptor function](#)

24 September 2018

Rita J. Valentino & Nora D. Volkow



[Opioid receptors: drivers to addiction?](#)

22 June 2018

Emmanuel Darcq & Brigitte Lina Kieffer

- Article
- Open Access
- Published: 03 May 2023

# Ligand and G-protein selectivity in the κ-opioid receptor


- Jianming Han,
- Jingying Zhang,
- Antonina L. Nazarova,
- Sarah M. Bernhard,

- Brian E. Krumm,
- Lei Zhao,
- Jordy Homing Lam,
- Vipin A. Rangari,
- Susruta Majumdar,
- David E. Nichols,
- Vsevolod Katritch,
- Peng Yuan,
- Jonathan F. Fay &
- …
- Tao Che

Show authors

*Nature* **volume 617**, pages 417–425 (2023)

- 7289 Accesses
- 126 Altmetric
- Metrics details

## Abstract

The κ-opioid receptor (KOR) represents a highly desirable therapeutic target for treating not only pain but also addiction and affective disorders[1]. However, the development of KOR analgesics has been hindered by the associated hallucinogenic side effects[2]. The initiation of KOR signalling requires the $G_{i/o}$-family proteins including the conventional ($G_{i1}$, $G_{i2}$, $G_{i3}$, $G_{oA}$ and $G_{oB}$) and nonconventional ($G_z$ and $G_g$) subtypes. How hallucinogens exert their actions through KOR and how KOR determines G-protein subtype selectivity are not well understood. Here we determined the active-state structures of KOR in a complex with multiple G-protein heterotrimers—$G_{i1}$, $G_{oA}$, $G_z$ and $G_g$—using cryo-electron microscopy. The KOR–G-protein complexes are bound to hallucinogenic salvinorins or highly selective KOR agonists. Comparisons of these structures reveal molecular determinants critical for KOR–G-protein interactions as well as

key elements governing $G_{i/o}$-family subtype selectivity and KOR ligand selectivity. Furthermore, the four G-protein subtypes display an intrinsically different binding affinity and allosteric activity on agonist binding at KOR. These results provide insights into the actions of opioids and G-protein-coupling specificity at KOR and establish a foundation to examine the therapeutic potential of pathway-selective agonists of KOR.

# Main

Opioid receptors are G-protein-coupled receptors (GPCRs) that have important roles in pain sensation. Almost all clinically used opioids act through the μ-opioid receptor (MOR). However, their use is associated with severe side effects, including a high potential for abuse, addiction and death due to respiratory depression in overdose[3]. The magnitude of these problems has led to a search for opioid alternatives for the treatment of pain and related conditions[4]. The activation of opioid receptors recruits downstream effectors, including heterotrimeric G proteins (including Gα, Gβ and Gγ subunits) and β-arrestins. Specifically, opioid receptors primarily couple to the $G\alpha_{i/o}$ family ($G_{i1}$, $G_{i2}$, $G_{i3}$, $G_{oA}$, $G_{oB}$, $G_z$ and gustducin ($G_g$)) (Extended Data Fig. 1a). Some of these subtypes can mediate non-overlapping signalling pathways depending on the GPCR involved[5,6,7,8]. Whether signalling through individual pathways has redundant roles or separately drives the therapeutic efficacy and side effects of opioids remains mostly unclear.

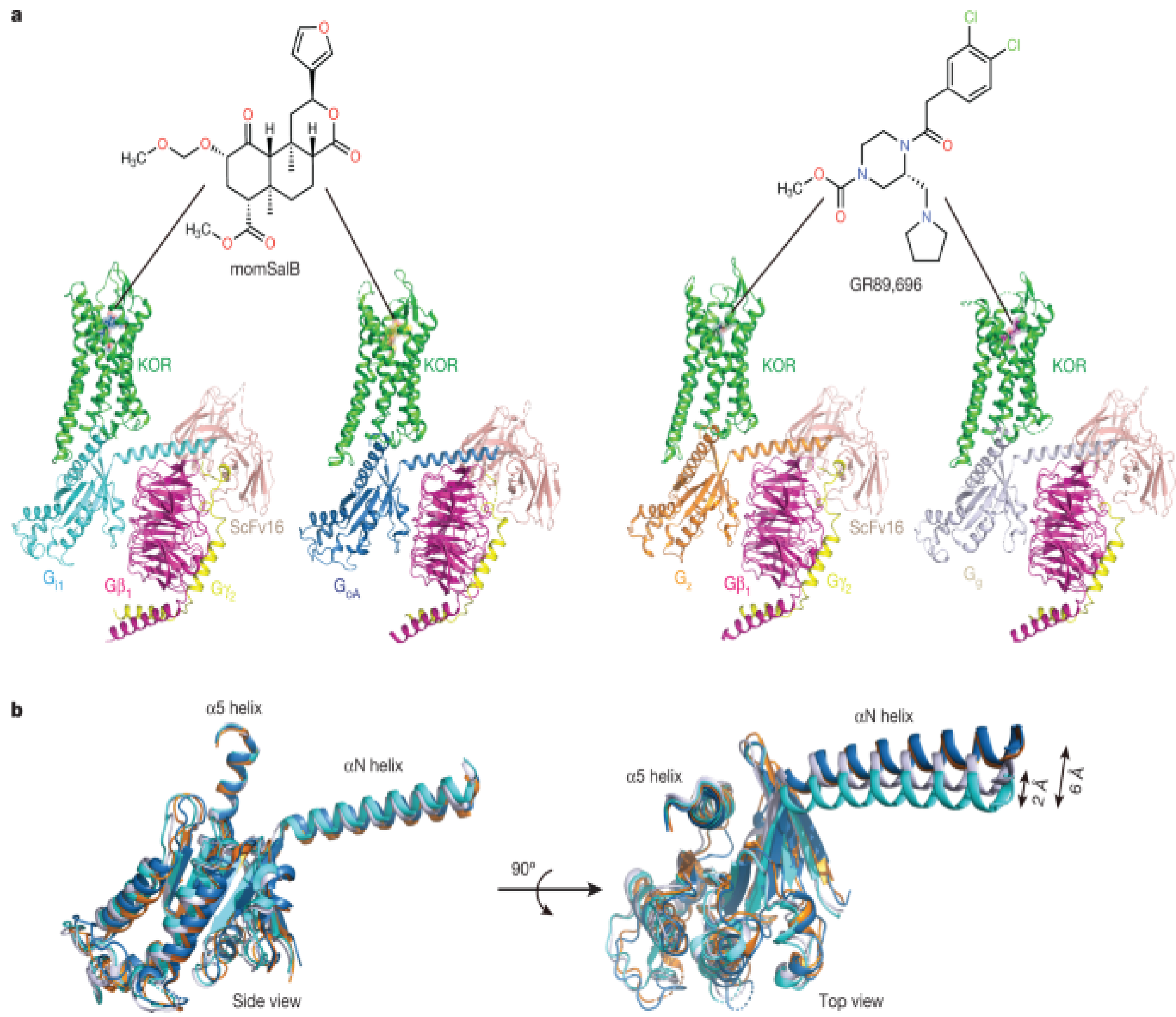
KOR is a highly desirable therapeutic target for treating not only pain but also addiction and affective disorders. KORs have gained increasing attention owing to their unique analgesic activity—they are predominantly expressed in pain-related neurons, and drugs that target KOR do not lead to addiction or cause death due to overdose as observed for MOR agonists[1]. The lack of rewarding/euphorigenic effects initially encouraged the development of KOR-agonist drugs as non-addictive analgesics[9]. Potent and selective KOR agonists have been developed, and these agonists produce effective peripheral and central analgesia. However, mood disorders such as dysphoria and psychotomimesis have been frequently observed as side effects of KOR agonists, which has limited their therapeutic application[2].

Here we determined the atomic structures of KOR in complex with different G-protein transducers and hallucinogenic ligands to help to elucidate the actions of opioids and the molecular basis for $G\alpha_{i/o}$ subtype selectivity.

## Overall structures of KOR–G-protein complexes

Although many efforts have been dedicated to the structural and molecular basis underlying the differences between G-protein and arrestin signalling, the roles of individual G-protein subtypes and the molecular determinants of subtype selectivity remain largely unclear. Sequence alignment of the seven $G_{i/o}$ subtypes suggests that they could be further grouped into four subclasses on the basis of sequence identity ($G_{i1}$, $G_{i2}$ and $G_{i3}$; $G_{oA}$ and $G_{oB}$; and $G_z$ and $G_g$) (Extended Data Fig. 1b). To further understand the role of KOR–G-protein coupling and signalling, we determined the structures of KOR in complexes with four representative $G_{i/o}$ subtypes ($G_{i1}$, $G_{oA}$, $G_z$ and $G_g$) at nominal resolutions of 2.71 Å, 2.82 Å, 2.65 Å, and 2.61 Å, respectively, using single-particle cryo-electron microscopy (cryo-EM; Fig. 1a, Supplementary Fig. 1 and Extended Data Table 1). In particular, KOR–$G_{i1}$ and KOR–$G_{oA}$ are bound to a psychotropic salvinorin analogue, methoxymethyl-salvinorin B (momSalB)[10]. However, cryo-EM experiments of KOR–$G_z$ or KOR–$G_g$ bound to momSalB yielded only low-resolution reconstructions (resolution of around 4.5–5 Å) that prevented the delineation of detailed molecular interactions. Thus, we leveraged another highly potent KOR agonist, GR89,696 (ref. [11]), to obtain high-resolution structures of KOR–$G_z$ and KOR–$G_g$.

**Fig. 1: Cryo-EM structures of KOR in complex with $G_{i/o}$ family subtypes.**




**a**, Cartoon representations of KOR–G-protein complexes. Structures of KOR–$G_{i1}$ and KOR–$G_{oA}$ are bound to momSalB. Structures of KOR–$G_z$ and $G_g$ are bound to GR89,696. **b**, Structural alignment of the four Gα subunits. Distances of movement from the N terminus are labelled.

The high-resolution maps of the four structures enabled unambiguous modelling of the agonist-bound heterotrimeric complexes (Supplementary Fig. 2). The overall differences between the four structures are subtle (root mean square deviations (r.m.s.d.) of 0.5 Å), with the exception of the Gα subunit in each complex (Fig. 1b). G-protein interactions with the receptor are canonically driven by the α5 and the N-terminal (αN) helices of the Gα subunit. The overlay of the four different G-protein subtypes showed that they adopt similar conformations in the α5 helix but differ in the extent of movement in the αN helix (Fig. 1b). In particular, relative to $G_{i1}$, both $G_{oA}$

and $G_z$ exhibit a 6 Å displacement in the αN helix, whereas $G_g$ has a smaller 2 Å displacement. Notably, alignments of the MOR–$G_{i1}$ structure[12] with KOR–$G_{i1}$ indicate that the αN helix of $G_{i1}$ in MOR displays a position that is distinct from that of KOR–$G_{i1}$, whereas the α5 helix shows an orientation and interaction pattern similar to those in KOR (Extended Data Fig. 1c).

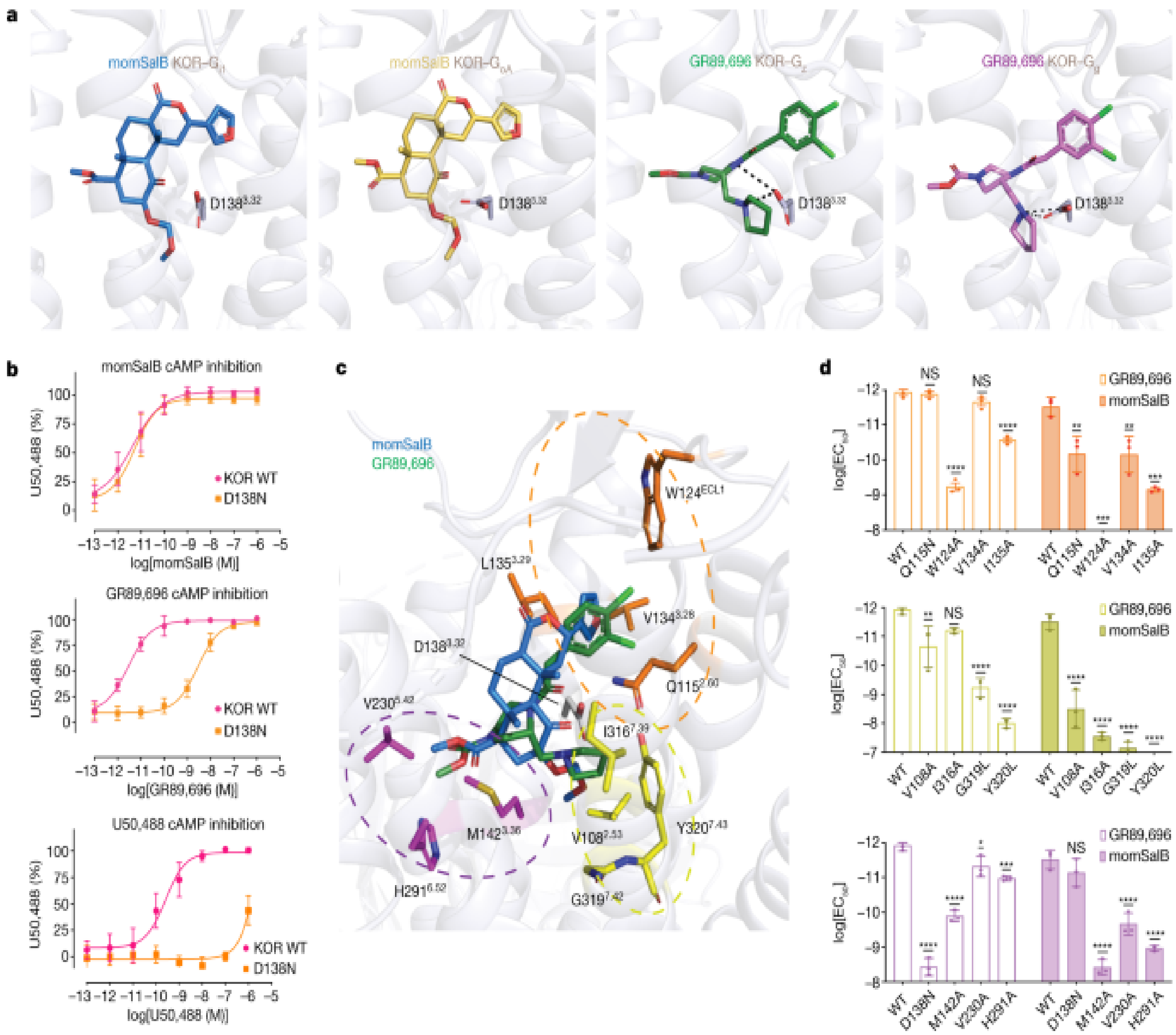
The overall structures of KOR in the $G_{i1/oA/z/g}$-bound states are similar to the previously reported nanobody-stabilized active conformation (KOR–Nb39)[13] (r.m.s.d., 0.8 Å) (Extended Data Fig. 1d). Notably, a comparison of these two structures reveals that the intracellular end of transmembrane helix 6 (TM6) in the KOR–$G_{i1}$ protein complex moves 2 Å closer to TM7. Nb39 stabilizing a different receptor conformation is further supported by its positive allosteric ability to enhance agonist binding affinity (Extended Data Fig. 1e). Another feature unique to G-protein-bound KOR is the presence of a well-defined intracellular loop 3 (ICL3) conformation that is absent in the Nb39-stabilized KOR, presumably due to its inherent flexibility (Extended Data Fig. 1d). Similar differences have also been captured between MOR–$G_{i1}$[12] or $\beta_2$AR–$G_s$[14] and their corresponding nanobody-stabilized active states[15,16], which further corroborate that a nanobody can stabilize a conformational state that mimics but is not identical to the G-protein-coupled state.

# Interactions of KOR with hallucinogenic salvinorins

KORs have a prominent role in the modulation of human perception. Salvinorins, such as salvinorin A (SalA)[17,18], are a group of naturally occurring hallucinogens with dissociative effects elicited by activating the central KORs. momSalB is a semi-synthetic analogue of SalA and displays similar in vivo pharmacology compared to SalA[19,20]. GR89,696 is a potent and long-lasting KOR agonist that produces antinociception and dysphoria but with unknown hallucinogenic properties[21]. Different binding poses of momSalB and GR89,696 were observed in the KOR orthosteric pocket. This is consistent with their divergent chemical structures—GR89,696 is an alkaloid (containing basic nitrogen atoms) and momSalB is a terpenoid

(lacking basic nitrogen atoms) (Fig. 1a). The pyrrolidine nitrogen atom in GR89,696, as well as many other ligands including KOR's endogenous dynorphin ligands[22], is essential for the binding to KOR and enables the ligand to act as a hydrogen-bond (H-bond) donor and forms a salt bridge with the carboxylate side chain of Asp138$^{3.32}$ in the binding pocket (where the superscript values indicate Ballesteros–Weinstein numbering for GPCRs[23]) (Fig. 2a). As salvinorin ligands (such as momSalB) lack the basic nitrogen atom, there are no attractive electrostatic interactions observed between the salvinorins and Asp138$^{3.32}$. Indeed, neither D138$^{3.32}$A nor D138$^{3.32}$N (the mutation in KOR DREADD[24]) showed detrimental effects in the binding affinity or agonistic potency of SalA, whereas both mutants abolished the interaction with endogenous dynorphin ligands[24,25,26]. The mutation D138$^{3.32}$N resulted in a significant loss of potency in U50,488 and GR89,696, but had minimal effects on momSalB (Fig. 2b). The side chain of Asp138$^{3.32}$ pointing to the methoxymethyl group of momSalB also explains an interesting observation that D138$^{3.32}$N could further enhance the binding affinity and potency of SalA and salvinorin B (SalB)[24], probably due to the switch from the unfavourable acceptor–acceptor interaction to attraction resulting from the new H-bond interactions between the side chain of mutated asparagine and methoxy oxygen of the ligand.

**Fig. 2: Ligand-specific interactions with KOR.**




**a**, The binding poses of momSalB and GR89,696 in their respective complex structures. The salt bridge or H-bond interactions in $G_z$- and $G_g$-coupled structures are shown as black dashed lines. This salt bridge or H-bond interaction is absent in momSalB-bound KOR. **b**, The highly conserved anchoring residue Asp138$^{3.32}$ has a different role in momSalB, GR89,696 or U50,488-mediated KOR activation. Data are normalized to the percentage of the reference agonist U50,488. Data are grouped data ± s.e.m. of $n = 3$ biological replicates. The full quantification parameters for this experiment are provided in Supplementary Table 1. **c**, Specific residues in the orthosteric pockets that interact with momSalB or GR89,696. Note that the I135$^{3.29}$L mutation was included in KOR structure constructs to increase the expression level. **d**, Mutagenesis screening of binding-pocket residues using G-protein-mediated cAMP inhibition assays. The effect on the potency

of momSalB or GR89,696 was quantified on the basis of the log[$EC_{50}$] values. Data are log[$EC_{50}$] ± s.e.m. of $n = 3$ biological replicates. Statistical significance for each mutant was calculated using one-way analysis of variance (ANOVA) with Dunnett's multiple-comparison test compared with the wild type (WT); *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$; NS, not significant. GR89,696: $P = 0.016$ (V230A), $P = 0.0008$ (H291A), $P = 0.1758$ (V134A), $P = 0.9814$ (Q115N), $P = 0.006$ (V108A), $P = 0.1165$ (I316A); momSalB: $P = 0.344$ (D138N), $P = 0.0009$ (W124A), $P = 0.0064$ (V134A), $P = 0.0068$ (Q115N), $P = 0.0002$ (I135A). The full quantification parameters for this experiment are provided in Supplementary Table 2.

Both momSalB and GR89,696 are highly selective and potent agonists at KOR (Fig. 2b and Extended Data Fig. 2a), making them ideal templates to investigate the molecular determinants for ligand selectivity and efficacy. Although the two agonists overlap in the orthosteric binding pocket of KOR, the core rings occupy different planes that are perpendicular to each other (Extended Data Fig. 2b). As a result, the subgroups of the two ligands form different interactions with residues in the corresponding subpockets (Fig. 2c). Mutations of the majority of residues in these subpockets reduced the agonist activity of momSalB or GR89,696, but with different amplitudes (for example, Val108$^{2.53}$, Gln115$^{2.60}$, Met142$^{3.36}$, Val230$^{5.42}$ or His291$^{6.52}$) (Fig. 2d and Extended Data Fig. 2c–e). The observation that the binding-pocket mutations have greater effects on momSalB-mediated cAMP inhibition than GR89,696 (for example, for H291$^{6.52}$A, Δlog[median effective concentration $(EC50)_{mutant-WT}$] = 2.23 ± 0.25 (momSalB) and 0.78 ± 0.27 (GR89,696)) is probably due to the lack of the anchoring interactions with Asp138$^{3.32}$, which makes salvinorins more sensitive to other residue contacts. The double mutation in KOR (for example, D138$^{3.32}$N and H291$^{6.52}$A, $pEC_{50} = 9.95 \pm 0.06$) displays a less deleterious effect on the potency of momSalB than H291$^{6.52}$A does ($pEC_{50} = 9.08 \pm 0.06$) alone (Extended Data Fig. 3a). A 2-fold to 2.5-fold improvement in potency was also observed from other mutations (Q115$^{2.60}$N or V230$^{5.42}$A) in combination with D138$^{3.32}$N when compared with the single mutation without D138$^{3.32}$N (Extended Data Fig. 3a). This effect might be specific to momSalB or salvinorin ligands as the double mutations (V230$^{5.42}$A/D138$^{3.32}$N or H291$^{6.52}$A/D138$^{3.32}$N) led to an inactive U50,488 or a further loss of potency for GR89,696-mediated cAMP

inhibition in V230$^{5.42}$A/D138$^{3.32}$N (9,120-fold) or H291$^{6.52}$A/D138$^{3.32}$N (1,288-fold) compared with the respective single mutation (Extended Data Fig. 3a). Another major difference between momSalB and GR89,696 is that momSalB mainly forms hydrophobic interactions with residues that specifically contribute to the high potency of momSalB, such as Val108$^{2.53}$, Val134$^{3.28}$, Val230$^{5.42}$ and Ile316$^{7.39}$. In particular, Val108$^{2.53}$ has also been indicated as a determinant of ligand selectivity between KOR and MOR or DOR, as the latter two opioid receptors have an Ala$^{2.53}$ at the corresponding position[27]. Another hydrophobic pocket formed by the side chains of Val108$^{2.53}$ and Tyr320$^{7.43}$ and the backbone of Gly319$^{7.42}$ appears to be a key determinant for agonist activity and receptor activation, as mutations of these residues significantly decreased or eliminated signal transduction of momSalB with a threefold reduction in its ligand-binding affinity (Fig. 2d and Extended Data Fig. 3b). Notably, the amplitude of interactions with residues in this hydrophobic pocket positively correlates with agonist potency because ligands with more extended interacting groups—such as SalB (-O-H), momSalB (-O-$CH_2$-O-$CH_3$) and ethoxymethyl SalB (-O-$CH_2$-O-$CH_2$-$CH_3$) (Extended Data Fig. 3c,d)—have displayed increased potency in activating KOR[19]. This subpocket at the bottom of the ligand-binding pocket acts as a potential allosteric connector to initiate the conformational changes of other microswitch motifs, including the sodium site, CW$^{6.48}$xxP and Pro$^{5.50}$-Ile$^{3.40}$-Phe$^{6.44}$ motifs[28,29].

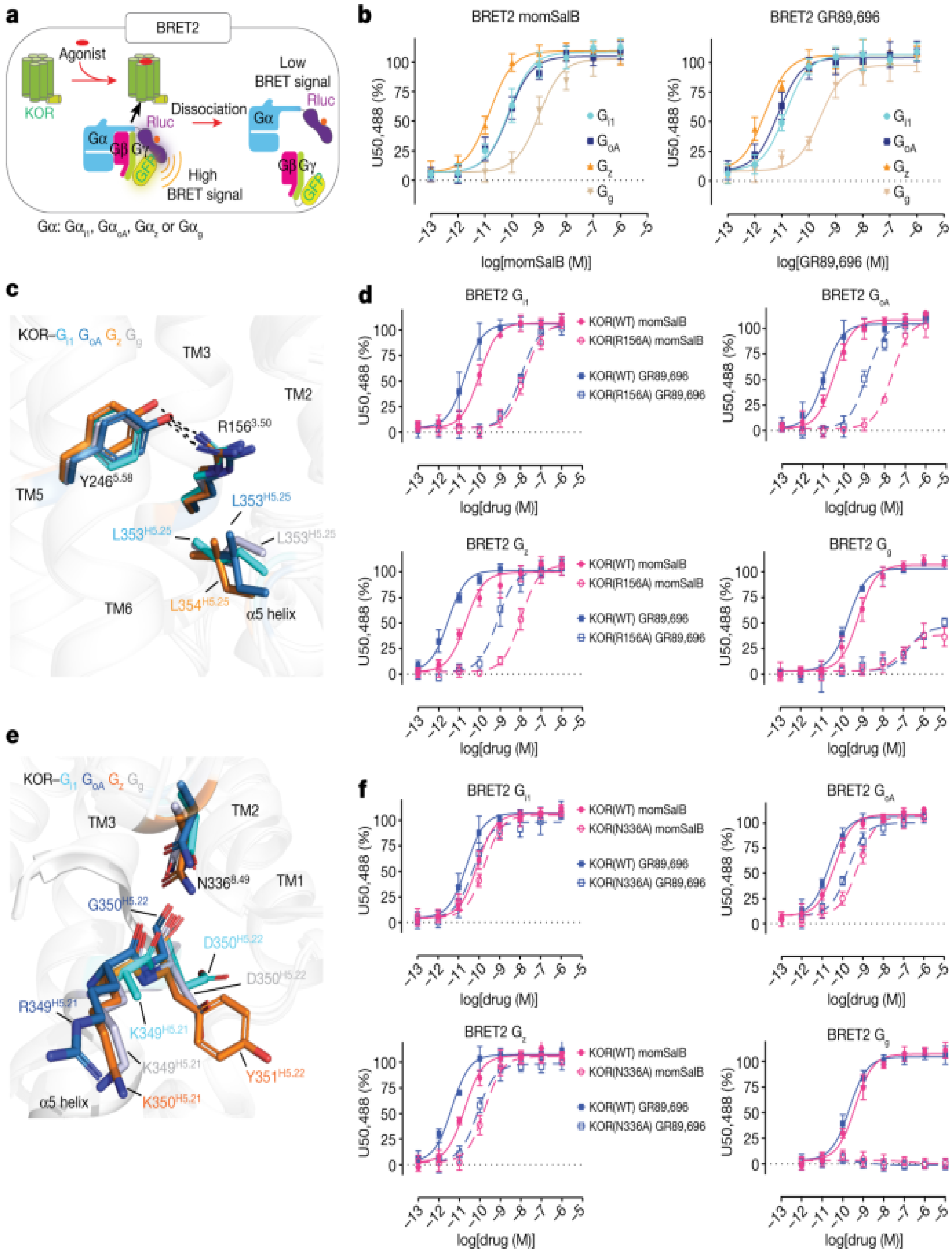
The overall binding pose of GR89,696 in KOR–$G_z$ is similar to that in KOR–$G_g$. One notable difference is that GR89,696 forms stronger salt-bridge interactions with Asp138$^{3.32}$ (2.9 and 3.4 Å) in KOR–$G_z$ than those in KOR–$G_g$ (3.5 and 3.9 Å) (Fig. 2a and Supplementary Fig. 3), which probably contributes to the higher potency in activating $G_z$ compared with $G_g$ (ref. [30]). Mapping the atomic distances between the ligand and receptor showed that GR89,696 makes closer contact with residues in the KOR–$G_z$ structure than in the KOR–$G_g$ structure (in terms of distance), whereas momSalB in KOR–$G_{i1}$ and KOR–$G_{oA}$ largely overlaps and displays similar strength (Supplementary Fig. 3). For example, GR89,696 in KOR–$G_z$ also forms H-bond interactions with Gln115$^{2.60}$ (2.8 Å) and His291$^{6.52}$ (3.3 Å),

and, in KOR–$G_g$, Gln115$^{2.60}$ (3.9 Å) and His291$^{6.52}$ (4.0 Å). This suggests that GR89,696 leads to more contractions of the ligand-binding pocket in the presence of $G_z$ compared with $G_g$.

## Structural basis of G-protein subtype selectivity

Similar to other opioid receptors, KOR exclusively couples to the $G_{i/o}$ family[30], including the canonical $G_{i/o}$ subtypes ($G_{i1}$, $G_{i2}$, $G_{i3}$, $G_{oA}$ and $G_{oB}$) and the noncanonical $G_z$ and $G_g$. Whereas $G_z$ is predominantly expressed in the central nervous system, $G_g$ is the endogenous transducer of taste receptors, such as the bitter taste receptor 2 (TAS2R). Mice expressing engineered KORs in bitter-receptor cells show a strong aversion to a designed KOR agonist (inert to endogenous wild-type KOR but active in engineered KOR)[31], suggesting that the KOR–$G_g$ interaction and signalling may also occur in vivo. Using bioluminescence resonance energy transfer (BRET)-based transducerome profiling (Fig. 3a), we confirmed that both momSalB and GR89,696 could activate all four G-protein subtypes, although with different potencies (Fig. 3b). The primary interaction sites in KOR bound to different $G_{i/o}$ subtypes involve nearly the entire intracellular regions of the receptor (ICL2, ICL3, TM3, TM5, TM6, TM7 and helix 8) and the αN and α5 helices of the Gα subunits (Extended Data Fig. 4). The key residues involved in KOR–G-protein interactions were mapped (Extended Data Fig. 4) and screened by alanine substitutions. In this section, we first report the effects of interface residues from the KOR side and then the residues from the Gα protein side.

**Fig. 3: Comparison of the receptor–G-protein-binding interface of the KOR–$G_{i1}$, KOR–$G_{oA}$, KOR–$G_z$ and KOR–$G_g$ complexes.**




**a**, Schematic of the BRET2 assay. **b**, momSalB- or GR89,696-mediated G-protein-subtype activation measured by BRET2. Data are grouped data ± s.e.m. of $n = 4$ biological replicates. The full quantification parameters for

this experiment are provided in Supplementary Table 3. **c**, Interactions of Arg156$^{3.50}$ in the Asp (D)-Arg (R)-Tyr (Y) motif with KOR and Gα. **d**, Mutagenesis analysis of Arg156$^{3.50}$ by BRET2. Data are grouped data ± s.e.m. of $n = 3$ biological replicates. The full quantification parameters for this experiment are provided in Supplementary Table 4. **e**, Interactions of Asn336$^{8.49}$ in different KOR–G-protein complexes. **f**, The N336$^{8.49}$ A mutation differentially affects KOR-mediated G-protein subtype activation. Data are global fit of grouped data ± s.e.m. of $n = 3$ independent biological replicates. The full quantification parameters for this experiment are provided in Supplementary Table 4.

Although the KOR conformations in each G-protein-bound structure are similar to each other, notable differences were observed for the KOR residues involved in receptor–G-protein interactions. Mutagenesis screening using G-protein-mediated cAMP inhibition assays suggested that almost all of the residues on the KOR side contribute to KOR–G-protein signalling (Extended Data Fig. 5a,b). That was further confirmed by the BRET-based transducer profiling, which showed that mutations of these residues on the intracellular KOR side decreased agonist-mediated $G_{i1}$, $G_{oA}$, $G_z$ and $G_g$ coupling (Extended Data Fig. 5c and Supplementary Fig. 4). Although most of the residues in the KOR interface affect the four G-protein couplings in a similar manner (Supplementary Fig. 4), some display subtype selectivity. Arg156$^{3.50}$ is a highly conserved residue in the classic Asp$^{3.49}$-Arg$^{3.50}$-Tyr$^{3.51}$ motif that has been implicated in having an important role in receptor activation and signal transduction (Fig. 3c). An 'ionic lock' has been frequently observed between Arg$^{3.50}$ and Glu$^{6.30}$ in class A GPCRs, keeping the receptor in an inactive state with TM3 and TM6 in close proximity. Thus, the breaking of this ionic lock is an important step towards the coupling of G proteins, as the TM6 movement away from TM3 is critical for penetration of the G-protein α5 helix into the cytoplasmic pocket[12]. The R156$^{3.50}$A mutation significantly reduced the potency of agonist-mediated activation by momSalB or GR89,696 (Fig. 3d). Furthermore, R156$^{3.50}$A specifically reduced the efficacy of agonist-mediated $G_g$ activation, momSalB (WT (106 ± 3%) versus R156$^{3.50}$A (38 ± 5%)) or GR89,696 (WT (103 ± 3%) versus R156$^{3.50}$A (46 ± 7%)) (Fig. 3d and Extended Data Fig. 5d). In the inactive-state KOR[32], the partially formed ionic lock is between

Arg156$^{3.50}$ and Thr273$^{6.34}$; in the fully active KOR–agonist–G-protein states, this interaction is broken due to the insertion of the α5 helix of Gα protein, leading to the release of the side chain of Arg156$^{3.50}$ to extend towards TM7 and form hydrophobic interactions with the second-to-last leucine (Leu353$^{H5.25}$ in $G\alpha_{i1}$, $G\alpha_{oA}$ or $G_{\alpha g}$; Leu354$^{H5.25}$ in $G\alpha_z$) of the Gα subunits (superscript notes for G proteins represent the CGN numbering system[33]) (Fig. 3c). This is further supported by the molecular dynamics simulations showing that the KOR-Arg156$^{3.50}$ can form hydrophobic interactions with Gα-L353$^{H5.25}$ or Leu354$^{H5.25}$ maintaining <4 Å distances with these side chains (Extended Data Fig. 6). In this extended conformation, the KOR-Arg156$^{3.50}$ guanidine group also forms a persistent H bond with Tyr246$^{5.58}$ observed in all four complexes (Fig. 3c and Extended Data Fig. 6). These data suggest that the KOR-Arg156$^{3.50}$ has an important role in KOR activation by directly interacting with the G proteins. A recent study also suggested that G proteins might need to precouple to the receptor and break the Arg156$^{3.50}$-mediated ionic interaction before agonist binding and signalling[34]. Other important interactions are formed by residue Asn336$^{8.49}$ in helix 8 of KOR, which engages different H-bond interactions with the backbone of the α5 helix in each Gα protein, such as Lys/Arg$^{H5.21}$ and Gly/Asp/Tyr$^{H5.22}$ (Fig. 3e and Extended Data Fig. 4). The molecular dynamics simulations also provide support for these interactions, suggesting dynamic patterns of switches between specific interaction pairs (Extended Data Fig. 7). The mutation N336$^{8.49}$A completely abolished KOR–$G_g$ coupling (for example, momSalB), and a 2-fold loss in potency in $G_{i1}$, 14-fold in $G_{oA}$ or 9-fold in $G_z$ coupling (Fig. 3f). Together with the effects observed from Arg156$^{3.50}$ and Asn336$^{8.49}$, these data indicate that these residues have differential roles in G-protein association, probably by engaging at different intermediate stages. The observation that several mutations have the largest effect on $G_g$ compared with the other $G_{i/o}$ subtypes suggests a non-canonical role of $G_g$ in KOR-mediated signalling.

Next, we examined the Gα subunit by mutating the non-conserved residues in the αN or α5 helix to alanine (Extended Data Fig. 8a). However, we did not observe significant changes in the potency of agonist-mediated G-protein activation in BRET2 assays (Extended Data Fig. 8b–f and

Supplementary Fig. 5). One exception is that C351$^{H5.23}$A in G$\alpha_{i1/oA/g}$ or I352$^{H5.23}$A in G$\alpha_z$ led to a significant decrease in potency for momSalB or GR89,696-mediated G-protein activation. This Ile352$^{H5.23}$ in G$\alpha_z$, compared with the corresponding Cys351$^{H5.23}$ in other G$_{i/o}$ subtypes, is known as the site that makes G$\alpha_z$ insensitive to pertussis toxin. The relative conformation of the α5 helix has been implicated as a key determinant between G$_s$ and G$_i$ specificity, in which the α5 helix adopts distinct positions and results in a larger outward movement of TM6 (13 Å in β$_2$AR–G$_s$ versus 9 Å in MOR–G$_{i1}$)[12,35,36]. The subtle differences in the α5 helix conformation of G$\alpha_{i1/oA/z/g}$ and the mutational evidence suggest that the G-protein-coupling specificity in the G$_{i/o}$ family is probably determined by a more complex and/or dynamic three-dimensional network interaction[37,38].

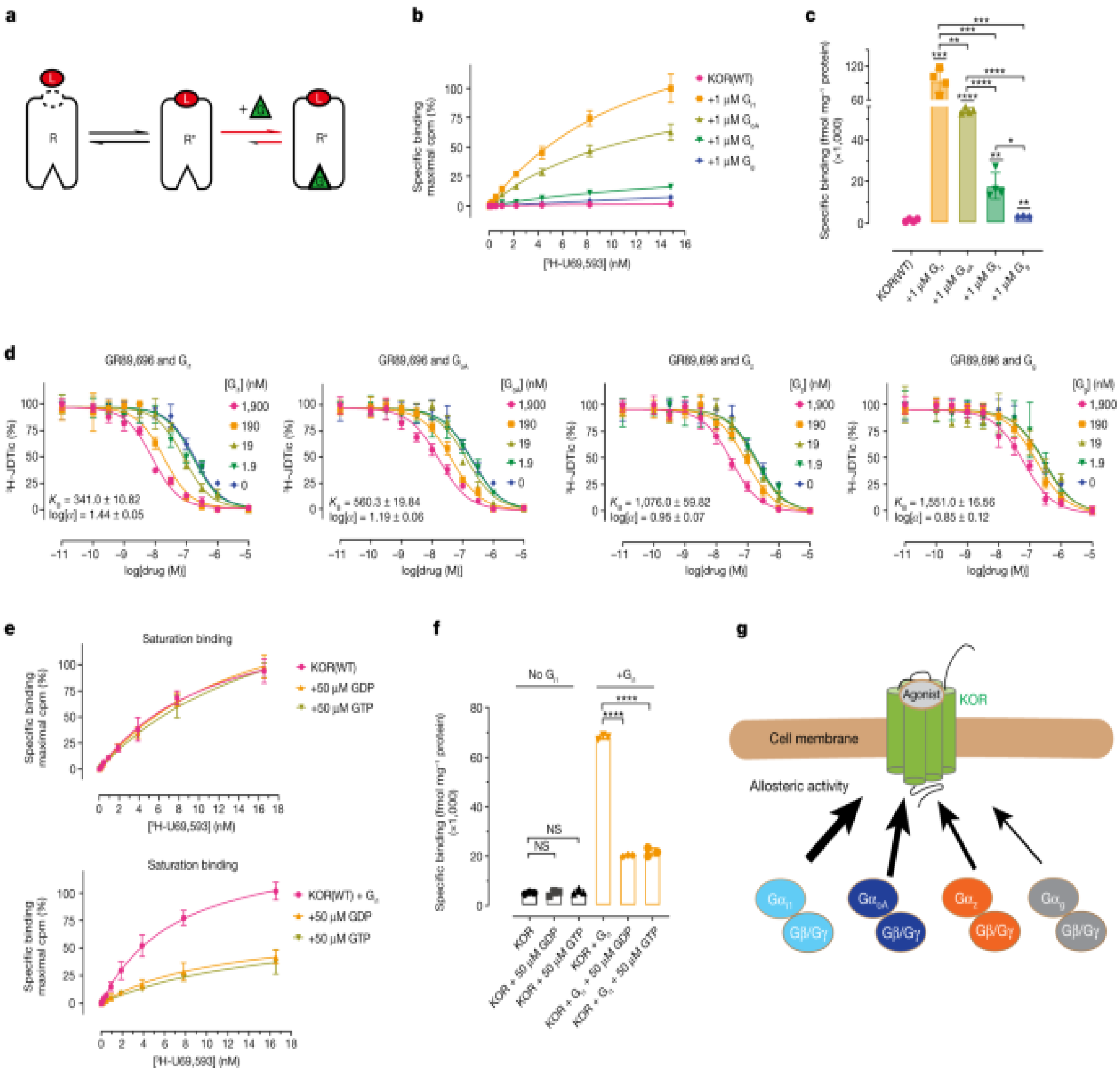
The overall interfaces of G$\alpha_{i1}$, G$\alpha_{oA}$, G$\alpha_z$ and G$\alpha_g$ with KOR are highly conserved (Extended Data Fig. 4), but there are critical differences in the α5 and αN helices. The major contacts made by G$\alpha_{oA}$ with KOR are through residues in the α5 helix (Extended Data Fig. 4b), whereas contacts made by G$\alpha_{i1}$, G$\alpha_z$ and G$\alpha_g$ involve regions in both the α5 and αN helices (Extended Data Fig. 4a,c,d). Similarly, the 5HT$_{1B}$R–G$_o$[39] interaction is also mediated solely by the α5 helix, but a structural comparison between KOR–G$\alpha_{oA}$ and 5HT$_{1B}$R–G$_o$ shows that the α5 helix in 5HT$_{1B}$R–G$_o$ tilts an additional 9°, leading to a larger 3 Å outward movement of TM6 (Extended Data Fig. 9a). Alignment of the cytoplasmic regions of KOR–G$\alpha_{i1}$, β$_2$AR–G$\alpha_s$ and 5HT$_{2A}$R–G$\alpha_q$ shows that the α5 helices are positioned differently. There are 6° and 12° tilts of the C-terminal end away from the plane of the membrane compared with G$\alpha_q$ and G$\alpha_s$, respectively, leading to different magnitudes of outward movement of TM6 (Extended Data Fig. 9b). As a result of intracellular conformational differences, the KOR–G$\alpha_{i1}$ forms an interface area of 1,219 Å$^2$ (Extended Data Fig. 4a), compared with a slightly larger area of β$_2$AR–G$\alpha_s$ (1,260 Å$^2$) and a much smaller area of 5HT$_{2A}$R–G$\alpha_q$ (1,077 Å$^2$) (Extended Data Fig. 9c). Whereas G$\alpha_{i1}$, G$\alpha_z$ and G$\alpha_g$ have similar interface areas (1,219, 1,262 and 1,221 Å$^2$, respectively) (Extended Data Fig. 4a,c,d), G$\alpha_{oA}$ has a much smaller area (1,096 Å$^2$) (Extended Data

Fig. 4b). Notably, the 822 Å$^2$ surface area of $G_o$ in contact with $5HT_{1B}R$[39] is closer to that of KOR and $G\alpha_{oA}$ compared with other G-protein subtypes, suggesting a shared mechanism between different GPCRs and the same G protein.

## Intrinsic differences in G-protein subtypes

GPCR signalling is transduced through the allosteric changes between the extracellular ligand pocket and the intracellular G-protein-binding pocket. Conformational changes induced by agonist binding can enhance the binding affinity of G-protein heterotrimers. Conversely, G protein acts as a positive allosteric modulator and further enhances agonist-binding affinity by stabilizing the ternary complex[40] formed by the receptor, ligand and G proteins (Fig. 4a). We next sought pharmacological evidence to test whether G-protein subtypes have intrinsic differences, including binding affinity and allosteric activity at KOR in the presence of agonists. On the basis of the ternary complex model[41], the high-affinity agonist-binding states should increase in the presence of G-protein heterotrimers, as the latter can stabilize the active-state receptor favouring agonist binding. We performed saturation binding assays to test the binding of agonist radioligand $^3$H-U69,593 to KOR in the presence of $G_{i1}$, $G_{oA}$, $G_z$ and $G_g$. Notably, the four G proteins display substantial differences in the allosteric enhancement of agonist binding (Fig. 4b,c). Compared with the wild type alone ($B_{max} = 1{,}350 \pm 116$), the high-affinity binding sites for $^3$H-U69,593 were increased 62-, 38-, 14- and 7-fold in the presence of $G_{i1}$ ($B_{max} = 84{,}324 \pm 4{,}214$), $G_{oA}$ ($B_{max} = 52{,}086 \pm 2{,}465$), $G_z$ ($B_{max} = 18{,}623 \pm 1{,}468$) and $G_g$ ($B_{max} = 9{,}866 \pm 3{,}493$), respectively. These data are consistent with the ternary model that at least two binding states predominate in the unliganded receptor[42]—a high-affinity (G-protein-coupled) and a low-affinity (G-protein-uncoupled) binding state. We also compared the wild-type G proteins with the engineered G proteins used in our structural determination and observed similar patterns of $B_{max}$ increases (Extended Data Fig. 10b).

**Fig. 4: The intrinsic differences of individual G-protein subtypes.**




**a**, Schematic of the GPCR–G-protein–ligand ternary model. G, G protein; L, agonist; R, receptor. **b**, KOR saturation binding reveals that G proteins potentiate agonist binding with different amplitudes. Data are global fit of grouped data ± s.e.m. from $n = 4$ independent biological replicates. **c**, Summary of $B_{max}$ values in the presence of G proteins. Statistical analysis between groups was performed using the unpaired two-tailed Student's *t*-tests; $P = 0.0001$ (KOR + $G_{i1}$ versus KOR), $P = 0.0021$ (KOR + $G_z$ versus KOR), $P = 0.0075$ (KOR + $G_g$ versus KOR), $P = 0.0081$ (KOR + $G_{i1}$ versus KOR + $G_{oA}$), $P = 0.0004$ (KOR + $G_{i1}$ versus KOR + Gz), $P = 0.0007$ (KOR + $G_{i1}$ versus KOR + $G_g$) and $P = 0.0116$ (KOR + $G_z$ versus KOR + $G_g$). **d**, Competition binding reveals that G proteins have a different binding affinity

and allosteric activity on KOR. Data are global fit of grouped data ± s.e.m. of $n = 3$ independent biological replicates. **e**, The effect of GDP or GTP on the allosteric activity of G proteins. Data are global fit of grouped data ± s.e.m. of $n = 3$ independent biological replicates. **f**, Summary of $B_{max}$ values in the presence or absence of GDP/GTP. Statistical analysis was performed using unpaired two-tailed Student's *t*-tests compared with the KOR or KOR + $G_{i1}$ group; $P = 0.9874$ (KOR + 50 μM GDP versus KOR), $P = 0.4147$ (KOR + 50 μM GTP versus KOR). **g**, A representative model of different allosteric activity of $G_{i/o}$ family subtypes ($G_{i1} > G_{oA} > G_z > G_g$). The full quantification parameters for the experiments in **b**, **d** and **e** are provided in Supplementary Tables 5–7, respectively.

The different magnitudes of $B_{max}$ increases among G-protein subtypes suggest that individual G proteins have different allosteric effects on ligand binding. To test this hypothesis, we next compared the cooperativity of the four G-protein subtypes by radioactive competition binding assays designed to quantify the G-protein binding affinity ($K_B$) and cooperativity ($\alpha$, G-protein cooperativity value, $\alpha > 1$ indicates positive effects in increasing agonist affinity)[43]. The G-protein-insensitive antagonist radioligand $^3$H-JDTic was used for the following series of experiments. The inhibition of $^3$H-JDTic binding at KOR by GR89,696 progressively improved as the concentration of G protein increased, indicating positive cooperativity between G proteins and agonist binding (Fig. 4d). The calculated $K_B$ and $\alpha$-cooperativity displayed a pattern similar to that observed in the saturation binding, in which $G_{i1}$ has the highest binding affinity and allosteric effects at KOR in the presence of agonists, and $G_g$ has the least (Fig. 4d). The different binding affinities could have a role in G-protein-subtype selectivity, as G-protein subtypes with higher affinity can outcompete other G-protein subtypes, depending on subtype abundance, especially in cells expressing several or all $G_{i/o}$ family subtypes.

Guanosine diphosphates (GDPs) or guanosine triphosphates (GTPs) are important regulators of GPCR–G-protein assembly and signalling[44]. We therefore examined whether the presence of GDP or GTP can affect the allosteric activity of G-protein subtypes. The specific binding of $^3$H-U69,593 was significantly reduced in the presence of GDP or GTP

compared with the nucleotide-free state (Fig. 4e,f). The four G proteins exhibited a uniform pattern, showing similar responses to GDP or GTP (Extended Data Fig. 10c). These nucleotide-specific effects are consistent with the results of single-molecule studies of the $\beta_2$AR–$G_s$ complex showing that the presence of GDP or GTP accelerates the dissociation of $\beta_2$AR and the $G_s$ heterotrimer[45], which is achieved through a sequential conformational change in the Gα subunit after the binding of GDP or GTP[44,46]. Note that the dominant negative $G_{i2}$ has been reported to abolish GTP binding and GTPase activity[47]; however, the engineered G proteins in this study appear to maintain the GTP binding affinity and GTP turnover activity, although at weaker levels compared with the wild type (Extended Data Fig. 10a).

# Discussion

After activation, KOR can interact with up to seven G proteins, the coupling of which determines the direction of ligand-induced signalling. The seven G-protein subtypes are highly homologous but not structurally or functionally identical. The binding of signal transducers is coupled with specific receptor conformational changes, such as the magnitude of TM6 displacement. Such conformational differences have been observed in GPCRs bound to the $G_s$, $G_i$ and $G_q$ families. However, analysis of the four structures of KOR in complex with different $G_{i/o}$ subtypes shows that the receptor adopts a similar conformation. In particular, the receptor conformations in KOR–$G_z$ and KOR–$G_g$ are nearly identical, although the bound agonist GR89,696 activates the two G proteins with a 100-fold difference in potency. This structural observation agrees with an original postulation that the cross-reactivity between receptors and G proteins speaks to the conservation of structure among the receptor-binding domains of the G proteins and the G-protein-binding domains of the receptors[48]. This conformational similarity, irrespective of transducer subtypes, has also been observed in other GPCRs engaging G proteins versus arrestins[49]. These subtle differences could be due to the limitation of the structures that reveal a well-resolved population of the nucleotide-free G-protein-bound conformational state of the receptor. These nucleotide-free states of Gα subunits ($G\alpha_{i1}$, $G\alpha_{oA}$, $G\alpha_z$ and $G\alpha_g$) tend to stabilize a specific

conformational state of KOR. In the absence of G proteins or in the presence of nucleotide-bound G proteins, the receptor can adopt dynamic conformations that are different from that captured by nucleotide-free Gα[45]. Other approaches, including nuclear magnetic resonance (NMR)[50] and molecular dynamics simulations[51], have identified dynamic conformational states in the intracellular regions of the receptor related to transducer couplings in the presence of GDP or GTP.

Different GPCR–G-protein interfaces have been proposed to contribute to the differential kinetics of G proteins during association and dissociation with the receptor[52,53]. Our pharmacological characterization of the KOR–G-protein interface identified key residues that have different roles in G-protein coupling. The complexes that we targeted in this study displayed varying interface areas dependent on the receptors and G proteins. However, time-resolved studies are needed in the future for the direct measurement of G-protein association and dissociation rates, especially for different $G_{i/o}$ subtypes, as the strength of the receptor–G-protein interface may be another factor that affects the coupling efficiency.

In the structures of KOR in complex with different G-protein subtypes, we also revealed the binding poses of two selective KOR agonists—momSalB and GR89,696. Although they occupied the same binding pocket, they adopted different conformations and interacting patterns, probably due to their unique chemical structures. GR89,696 displays stronger interactions in KOR–$G_z$ than in KOR–$G_g$, which may contribute to its higher potency observed in the BRET2 assay. Owing to the unique scaffold and pharmacology of salvinorin ligands, extensive studies have been conducted to elucidate their binding and function[13,25,26]. Several residues or motifs in KOR (for example, Val108$^{2.53}$, His291$^{6.52}$, Ile316$^{7.39}$) have been identified that are important for salvinorin's agonism, which can now be explained by their direct interactions with momSalB. Notably, using multiple structural templates, previous salvinorin docking suggested different binding poses[13,25,26], and our structures now provide direct evidence of how momSalB sits in the binding pocket of KOR.

We also observed allosteric differences among these highly conserved G-protein subtypes. As positive allosteric modulators, the four representative G

proteins display distinct allosteric activity in potentiating agonist binding ($G_{i1} > G_{oA} > G_z > G_g$) (Fig. 4g). This is consistent with measurements of the binding affinity of different G proteins ($G_{i1} > G_{oA} > G_z > G_g$) at KOR. These intrinsic differences in G proteins, including the binding affinity and coupling efficiency, add pharmacological evidence to determinants for G-protein-subtype selectivity. Our structural observations from different G proteins in complex with KOR show that the $G_{i/o}$ family subtypes share a highly conserved mechanism in interacting with KOR, but that each maintains pharmacological differences. Considering that many GPCRs can couple to different G-protein families, such as the $\beta_2$AR coupling with both $G_s$ and $G_{i/o}$ (refs. [38,54]), whether $\beta_2$AR displays differential binding affinities with $G_s$ and $G_{i/o}$ may help to explain its G-protein coupling specificity.

Furthermore, the allosteric activity of G proteins can be regulated by GDPs or GTPs that decouple G proteins from the receptor. It is known that nucleotide-specific conformations exist between nucleotide-free G proteins and GDP- or GTP-bound G proteins. When comparing the crystal structure of the uncoupled GDP-bound $G_{i1}$ heterotrimer[55] and nucleotide-free (KOR–$G_{i1}$, $G_{oA}$, $G_z$ and $G_g$) heterotrimers (Extended Data Fig. 11a), several conformational displacements are noted (Extended Data Fig. 11b). The activated receptor engages the C terminus of the α5 helix of $G\alpha_{i1}$, which undergoes an upward helical extension (8.6 Å) into the receptor core (Extended Data Fig. 11c) compared with the uncoupled G protein structure. The insertion of the α5 helix into the transmembrane helical bundle of the receptor has the following two consequences. First, the loop connecting the α5 helix and β6 strand moves outward 5 Å. Second, the movement of the α5 helix disrupts the original hydrophobic interactions between the α5 and α1 helices, leading to a displacement of the P loop. Both the P loop translocation and loss of coordination with GDPs are necessary for GDP release[44,45]. In agreement with the ternary complex model, our saturation binding data show that GDPs or GTPs act as negative regulators of agonist binding kinetics in the presence of G proteins. One limitation of this study is that we used an in vitro overexpressed system with engineered receptors and G proteins to measure ligand activity, which cannot be extended to in vivo without further experiments.

In summary, we have elucidated the molecular interaction details between highly conserved $G_{i/o}$ subtypes and KOR using cryo-EM-derived atomic models. We have also examined the structural determinants of ligand selectivity and efficacy at KOR. Using structural pharmacology analysis, we revealed the intrinsic differences between these previously under-represented $G_{i/o}$ subtypes and demonstrated that subtype selectivity is probably a combinational result of receptor conformational dynamics, the binding affinity of G proteins and cooperativity between agonist binding and G-protein coupling. Such findings are important both in understanding GPCR-mediated signalling and in the generation of new research tools and therapeutics based on the potential of G-protein-selective agonists.

# Methods

## Generation of constructs for cryo-EM

For the human KOR, we used a construct the same as the previously determined active-state KOR[13]. In brief, the construct (1) lacks N-terminal residues 1–53; (2) lacks C-terminal residues 359–380; (3) contains Met1–Leu106 of the thermostabilized apocytochrome b562 RIL (BRIL) from *E. coli* (M7W, H102I, R106L) in place of receptor N terminus residues Met1–His53. This N-terminal Bril will be removed using a PreScission cleavage site in the end. The single chain Fab scFv16 has the same sequence as previously reported[56]. A 6×His tag was added to the C-terminal scFv16 sequence with a PreScission cleavage site inserted between. For the G-protein heterotrimers, individual G-protein constructs ($G_{i1}$, $G_{oA}$, $G_z$ and $G_g$) were engineered (labelled as dominant negative, DN)[47] for the binding of scFv16, and then subcloned into a designed vector that co-expresses the $G\beta_1$ and $G\gamma_2$. Further modifications were made to enable a stable complex between KOR, G-protein heterotrimer and scFv16. Specifically, $G_{i1}$(DN) includes S47N, E245A, G203A and A326S. $G_{oA}$(DN) includes C3S, S47N, G204A, E246A, A326S and M249K. For $G_z$(DN), the N-terminal sequence was replaced with the $G_{i2}$ sequence to allow for better interaction with scFv16; other mutations include S47N, G204A, E246A, R249K, N262D and A327S. For $G_g$(DN), the N-terminal sequence was replaced with the $G_{i2}$

sequence; other mutations include S47A, G203A, E245A, H248K, T261D, A326S and N251D.

## Expression of KOR–G-protein–scFv16 complex

The Bac-to-Bac Baculovirus Expression System was applied to generate high-quality recombinant baculovirus (>$10^{-9}$ viral particles per ml) for protein expression (KOR, G-protein heterotrimers and scFv16). For the expression of KOR–G–scFv16 protein complex, each heterotrimeric G protein, including Gα ($G_{i1}$, $G_{oA}$, $G_z$ or $G_g$), $G\beta_1$ and $G\gamma_2$ was coexpressed with KOR and scFv16, respectively, by infection of *Spodoptera frugiperda* Sf9 cells at a cell density of $2.5 \times 10^6$ cells per ml in ESF921 medium (Expression System) with the P1 baculovirus at a multiplicity of infection (MOI) ratio of 2:2:0.5. Cells were collected by centrifugation (125 rpm at 27 °C) for 48 h after infection, washed with HN buffer (25 mM HEPES pH 7.4, 100 mM NaCl), and stored at −80 °C for future purification.

## Purification of the KOR–G-protein–scFv16 complex

The compounds used in this study—(−)-U50,488 (0496) and GR89,696 (1483)—were purchased from Tocris. momSalB was synthesized by a method described previously[57]. After purification by silica gel column chromatography, momSalB was a single spot on TLC (silica, 20% ethyl acetate, dichloromethane) with an $R_f$ of 0.49. An NMR spectrum of momSalB was collected to confirm the chemical identity (Supplementary Fig. 6), which is consistent with the expected spectrum reported previously[58].

We thawed the cell pellet and incubated it in buffer containing 20 mM HEPES pH 7.5, 50 mM NaCl, 1 mM $MgCl_2$, 2.5 units Apyrase (NEB), 10 μM agonist (final concentration) and protease inhibitors (500 mM AEBSF, 1 mM E-64, 1 mM leupeptin, 150 nM aprotinin) for 1.5 h at room temperature. We then collected the membrane by centrifugation at 25,000 rpm for 30 min at 4 °C. The membrane was solubilized in buffer (40 mM HEPES pH 7.5, 100 mM NaCl, 5% (w/v) glycerol, 0.6% (w/v) lauryl maltose neopentyl glycol (LMNG), 0.06% (w/v) cholesteryl hemisuccinate (CHS), 10 μM

agonist and protease inhibitors) with 200 μg scFv16 in the cold room. After 5 h, the supernatant was collected by centrifugation at 30,000 rpm for 30 min at 4 °C and incubated with 1 ml TALON IMAC resin (Clontech) and 20 mM imidazole overnight in the cold room. The next day, the resin was collected and washed with 10 ml buffer containing 20 mM HEPES pH 7.5, 100 mM NaCl, 30 mM imidazole, 0.01% (w/v) LMNG, 0.001% (w/v) CHS, 5% glycerol and 5 μM agonist. The protein was then eluted with the same buffer supplemented with 300 mM imidazole, concentrated and further purified by size-exclusion chromatography on the Superdex 200 increase 10/300 column (GE healthcare), which was pre-equilibrated with 20 mM HEPES pH 7.5, 100 mM NaCl, 100 μM TCEP, 0.00075% (w/v) LMNG, 0.00025% (w/v) glyco-diosgenin (GDN) and 0.00075% (w/v) CHS, 1 μM agonist. Peak fractions were collected, concentrated and incubated with PNGase F (NEB), PreScission protease (GenScript) to remove the potential glycosylation and N-terminal His–BRIL, respectively, and 100 μg scFv16 at 4 °C overnight. The next day, cleaved His–BRIL and protein, uncleaved protein and proteases were separated by the same procedure as described above. Peak fractions were concentrated to 3–5 mg $ml^{-1}$ for electron microscopy analysis. Four KOR–G-protein–scFv16 complexes were purified according to the same procedure except that different agonists were used.

## Expression and purification of scFv16 protein

The scFv16 protein was expressed by infection of Sf9 cells at a cell density of $2.5 \times 10^6$ cells per ml in ESF921 medium (Expression System) with the P1 baculovirus at an MOI of 2. After 96 h, the cell culture medium containing secreted scFv16 protein was collected by centrifugation at 4,000 rpm for 15 min. The pH of the supernatant was adjusted to 7.5 by addition of Tris-base power. Chelating agents were quenched by the addition of 1 mM nickel chloride and 5 mM calcium chloride and incubation with stirring for 1 h at room temperature and 5 h in the cold room. We removed the precipitates by centrifugation and the resultant supernatant was further cleaned with 0.45 μm filter paper, and incubated with 2 ml Ni-NTA resin and 10 mM imidazole overnight in the cold room. The Ni-NTA resin was washed the next day with 20 ml buffer (20 mM HEPES pH 7.5, 100 mM NaCl, 0.00075% (w/v) LMNG, 0.000075% (w/v) CHS, 0.00025% (w/v) GDN, 20 mM imidazole). The protein was eluted with the same buffer supplemented

with 300 mM imidazole, concentrated and further purified on the Superdex 200 increase 10/300 column. Monomeric fractions were pooled, concentrated, flash-frozen in liquid nitrogen and stored at −80 °C until future use.

## Expression and purification of heterotrimeric G proteins

The expression of heterotrimeric G protein was achieved by infection of Sf9 cells at a cell density of $2.5 \times 10^6$ cells per ml in ESF921 medium (Expression System) with the P1 baculovirus at an MOI of 2. After 48 h, cells were collected and lysed in buffer containing 200 mM NaCl, 40 mM HEPES pH 7.5, 0.2% Triton X-100, 5% glycerol, 3 mM β-me and protease inhibitors. The supernatant was isolated by centrifugation at 40,000 rpm for 50 min and incubated with 1 ml Ni-NTA resin and 20 mM imidazole overnight at 4 °C. The resin was collected the next day and washed with 20 ml buffer containing 100 mM NaCl, 20 mM HEPES pH 7.5, 5% glycerol, 20 mM imidazole and 3 mM β-me. The protein was then eluted with elution buffer (300 mM NaCl, 20 mM HEPES pH 7.5, 5% glycerol, 3 mM β-me and 300 mM imidazole), concentrated and further purified on the Superdex 200 increase 10/300 column, which was pre-equilibrated with buffer the same as the elution buffer except without the imidazole. The peak fractions were concentrated, flash-frozen in liquid nitrogen and stored at −80 °C for future binding assays.

## Cryo-EM data collection and 3D reconstruction

The purified samples (3–4 μl) were applied to glow-discharged 300-mesh Au grids (Quantifoil R1.2/1.3) individually and vitrified using a Vitrobot mark IV (Thermo Fisher Scientific). Cryo-EM imaging was performed on the Talos Artica system operated at 200 kV at a nominal magnification of ×45,000 using a Gatan K3 direct electron detector at a physical pixel size of 0.88 Å. Each stack video was recorded for 2 to 2.7 s in 60 frames at a dose rate of about 15 $e^{-}\ px^{-1}\ s^{-1}$, leading to a total exposure dose indicated in Extended Data Table 1. Videos were collected automatically with SerialEM[59] using an optimized multishot array procedure[60].

Dose-fractioned image stacks were processed for beam-induced motion correction followed by contrast transfer function estimation. Particles were selected using Blob particle picker, extracted from the micrograph and then used for 2D classification and 3D classification followed by non-uniform refinement. All of these steps were performed in cryoSPARC[61,62].

## Model building and refinement

Maps from cryoSPARC were used for map building, refinement and subsequent structural interpretation. The dominant-negative $G_{i1}$ trimer model and scFv16 model were adapted from the cryo-EM structure of the MRGPRX2–$G_{i1}$ complex (Protein Data Bank (PDB): 7S8M)[63]. $G_{oA}$, $G_z$ and $G_g$ trimer models were built from the $G_{i1}$ trimer model, followed by mutating the non-conserved residues back to the wild-type $G_{oA}$, $G_z$ and $G_g$. The receptor KOR model was taken from the active-state KOR–Nb39 structure (PDB: 6B73)[13]. The receptor, G proteins and scFv16 were docked into the cryo-EM map using Chimera[64]. The complex models (KOR–G-protein–scFv16) were manually built in Coot[65], followed by several rounds of real-space refinement using Phenix[66]. The model statistics were validated using Molprobity[67]. Structural figures were prepared using Chimera or PyMol (https://pymol.org/2/).

## cAMP inhibition assay

For the KOR–$G\alpha_i$-mediated cAMP inhibition assay, HEK293T (ATCC CRL-11268) cells were co-transfected with human KOR or various mutants along with a split-luciferase-based cAMP biosensor (GloSensor, Promega) at a 1:1 ratio (KOR:GloSensor). After 16 h, the transfected cells were plated into poly-l-lysine-coated 96-well white clear-bottom cell culture plates with DMEM + 1% dialysed FBS at a density of 40,000–50,000 cells per 200 µl per well and incubated at 37 °C with 5% $CO_2$ overnight. The next day, 3× drug solutions were prepared in fresh drug buffer (20 mM HEPES, 1× HBSS, 0.3% bovine serum albumin (BSA), pH 7.4). The plates were decanted the next day and received 40 µl per well of drug buffer (20 mM HEPES, 1× HBSS, pH 7.4) followed by addition of 20 µl of 3× drug solutions for 15 min in the dark at room temperature. Cells then received 20

μl luciferin (4 mM final concentration) supplemented with isoproterenol (300 nM final concentration), stimulating the production of endogenous cAMP through $\beta_2$ adrenergic $G_s$ activation, and incubated in the dark at room temperature. After 15 min, luminescence intensity was quantified using the Mithras LB 940 multimode microplate reader (Berthold Technologies). Data were plotted as a function of drug concentration, normalized to percentage U50,488 stimulation, and analysed using log (agonist) versus response in GraphPad Prism (v.9.3.1).

## BRET2 assay

To measure the agonist-stimulated G-protein (wild type and mutants) activation by KOR and various mutants, a BRET2-based cell assay was used. Specifically, four plasmids (KOR, Gα, Gβ, Gγ) were used, in which each Gα is tagged with a luciferase (Rluc8) and Gγ is tagged with an N-terminal GFP. Specifically, the $G\alpha_{i1}/G\beta_3/G\gamma_9$, $G\alpha_{oA}/G\beta_3/G\gamma_8$, $G\alpha_z/G\beta_3/G\gamma_1$ and $G\alpha_g/G\beta_3/G\gamma_1$ combinations were used for BRET2 $G_{i1}$, $G_{oA}$, $G_z$ and $G_g$ experiments, respectively. Detailed information of the GFP-Gγ and Gα-Rluc8 constructs was described previously[30]. HEK293T cells were then transfected with the four plasmids (KOR, Gα-Rluc8, Gβ, Gγ–GFP) using a 1:5:5:5 DNA ratio of receptor:Gα-RLuc8:Gβ:Gγ-GFP2 (100 ng receptor, 500 ng Gα–RLuc8, Gβ and Gγ–GFP2 for 10 cm dishes). Transit 2020 (Mirus Biosciences) was used to complex the DNA at a ratio of 2 μl Transit per μg DNA in Opti-MEM (Gibco-Thermo Fisher Scientific). Then, 16 h after transfection, cells were plated in poly-l-lysine-coated 96-well white clear-bottom plates in plating medium (DMEM + 1% dialysed FBS) at a density of 40,000–50,000 cells in 200 μl per well and incubated overnight. The next day, the plates were decanted and washed once with 60 μl drug buffer (20 mM HEPES, 1× HBSS, pH 7.4) and then 60 μl drug buffer containing coelenterazine 400a (Nanolight Technology) at a final concentration of 5 μM was added to each well. After 5 min for substrate diffusion, 30 μl 3× drug solutions in fresh drug buffer (20 mM HEPES, 1× HBSS, 0.3% BSA, pH 7.4) was added to each well and incubated for an additional 5 min. Finally, the plates were read on the Mithras LB 940 multimode microplate reader (Berthold Technologies) with 400 nm (RLuc8-coelenterazine 400a) and 510 nm (GFP2) emission filters for 1 s per well. The GFP to Rluc8 ratio was calculated, plotted as a function of drug concentration, normalized to

percentage U50,488 stimulation and analysed using log (agonist) vs response in GraphPad Prism (v.9.3.1).

## Radioligand-binding assay

Saturation binding assays were performed using the construct BRIL-wt-$KOR_{(54–368)}$ reconstituted into nanodiscs comprised of KOR, spMSP1D1 and lipid mixture (POPC:POPE:POPG = 3:1:1) at a molar ratio of 1:3:100. Binding assays were set-up in 96-well plates in standard binding buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 0.1 mM EDTA, 0.1% BSA, pH 7.4) at room temperature. Saturation binding assays with 0.1–20 nM $^3$H-U69,593 in the standard binding buffer were performed to determine the equilibrium dissociation constant ($K_d$) and $B_{max}$. To determine the effects of G proteins on $^3$H-U69,593 binding, each G protein (final concentration 1 μM) was incubated with $^3$H-U69,593 and homogenous membrane fractions for 3.5 h at room temperature. Data were analysed using GraphPad Prism (v.9.3.1) using a one-site model.

For the competitive binding assay, $^3$H-JDTic (0.68 nM), homogenous membrane fractions expressing KOR and 3× GR89,696 solutions were incubated in 96-well plates in standard binding buffer in the absence or presence of four G proteins in various concentrations (final concentration: 1,900 nM, 190 nM, 19 nM, 1.9 nM, 0 nM) for 3.5 h at room temperature in the dark, and then terminated by rapid vacuum filtration onto chilled 0.3% PEI-soaked GF/A filters followed by three quick washes with cold wash buffer (50 mM Tris-HCl, pH 7.4) and read. Results (with or without normalization) were analysed using GraphPad Prism (v.9.3.1) using one-site or allosteric $IC_{50}$ shift models.

## Cell-surface expression studies

The cell-surface expression levels of wild-type KOR and its mutants were measured using an enzyme-linked immunosorbent assay (ELISA). In brief, HEK293T (ATCC CRL-11268) cells were transiently transfected with wild-type KOR and KOR mutant DNA at the same quantity. After 24 h, cells were plated in poly-l-lysine-coated 96-well white clear-bottom plates in

plating medium (DMEM + 1% dialysed FBS) at a density of 40,000–50,000 cells in 200 μl per well and incubated overnight. The next day, plates were decanted and fixed with 4% (w/v) paraformaldehyde for 10 min at room temperature. Cells were then washed twice with 1× phosphate-buffered saline (PBS) (pH 7.4) and blocked by 1× PBS containing 0.5% (w/v) non-fat milk for at least 30 min at room temperature followed by incubation with anti-Flag (M2)–horseradish peroxidase-conjugated antibodies (Sigma-Aldrich, A8592) diluted 1:20,000 in the same buffer for 1 h at room temperature. After washing three times with 1× PBS, 1-Step Ultra-TMB ELISA substrate (Thermo Fisher Scientific, 34028) was added to the plates and the plates were incubated at 37 °C for 15–30 min and terminated by addition of 1 M sulfuric acid ($H_2SO_4$) stop solution. Finally, the plates were read at a wavelength of 450 nm using the BioTek Luminescence reader. The data were analysed using GraphPad Prism (v.9.3.1).

## G-protein expression studies

To measure the expression levels of four wild-type G proteins and their mutants, HEK293T (ATCC CRL-11268) cells were transiently transfected with the same quantity of wild-type and mutant G proteins DNA. After 16 h, cells were plated in poly-l-lysine-coated 96-well white clear-bottom plates in plating medium (DMEM + 1% dialysed FBS) at a density of 40,000–50,000 cells in 200 μl per well and incubated overnight. The next day, the plates were decanted and washed once with 60 μl drug buffer (20 mM HEPES, 1× HBSS, pH 7.4), then 60 μl drug buffer containing coelenterazine 400a (Nanolight Technology) at a final concentration of 5 μM was added to each well. After 5 min for substrate diffusion, plates were read in a Mithras LB 940 multimode microplate reader (Berthold Technologies) with 400 nm (RLuc8-coelenterazine 400a) and 510 nm (GFP2) emission filters for 1 s per well. The Rluc8 values represented the G-protein expression levels and were plotted in the GraphPad Prism (v.9.3.1).

## GTP turnover assay

Analysis of GTPase activity of G proteins ($G_{i1}$, $G_{oA}$, $G_z$, $G_g$) was performed by using a modified protocol of the GTPase-Glo assay (Promega). G proteins were serially (1:1) diluted into various concentrations with a buffer

of 300 mM NaCl, 20 mM HEPES pH 7.5 and 1 mM DTT, and 5 µl was dispensed into each well of a 384-well plate. The reaction was initiated by adding 5 µl 1 µM GTP solution to 5 µl G proteins. After incubation for 90 min at room temperature, 10 µl reconstituted GTPase-Glo reagent was added to the sample and incubated for 30 min at room temperature. Luminescence was measured after addition of 20 µl detection reagent and incubation for 10 min at room temperature using the Mithras LB 940 multimode microplate reader (Berthold Technologies). The data were analysed using GraphPad Prism (v.9.3.1).

## Molecular dynamics simulations

The Gromacs simulation engine (v.2020.3)[68] was used to run all molecular dynamics simulations under the Charmm36 force-field topologies and parameters[69,70]. Charmm force-field parameters and topologies for the ligands momSalB and GR89,696 were generated using Charmm-GUI's Ligand Reader & Modeller tool[70]. The loop grafting and optimization for modelling missing side chains and loops was performed in the ICM-Pro (v.3.9-2b) molecular modelling and drug discovery suite (Molsoft)[71]. The structurally conserved helix-8 (Hx8) amphipathic helical motifs in KOR were modelled using human antagonist-bound KOR (PDB: 4DJH)[26] as the template structure. The lobe in $G_{i1}$, $G_{oA}$, $G_z$ and $G_g$ proteins was modelled using a human agonist-bound CB2–$G_i$ structure (PDB: 6PT0)[72]. Structure regularization and torsion profile scanning were performed using ICMFF force field[73]. The GR89,696-bound structures of KOR complexes with $G_z$ and $G_g$ proteins as well as momSalB-bound KOR with $G_{i1}$ and $G_{oA}$ proteins were then uploaded to the Charmm-GUI webserver[69], where the starting membrane coordinates were determined by the PPM[74] server using the Charmm-GUI interface. The complexes were then embedded in a lipid bilayer composed of 1,2-dipalmitoyl-sn-glycero-3-phosphatidylcholine (DPPC), 1,2-dioleoyl-sn-glycero-3-phosphatidylcholine (DOPC) and cholesterol (CHL1) following the recommended ratio of 0.55:0.15:0.30, respectively[75]. The GR89,696-bound KOR complex with $G_z$ contained 330 DPPC, 90 DOPC and 180 CHL1 lipids, 64,400 water molecules, and 178 sodium and 176 chloride ions. The GR89,696-bound KOR complex with $G_g$

contained 330 DPPC, 90 DOPC and 180 CHL1 lipids, 64,227 water molecules, and 184 sodium and 175 chloride ions. The momSalB-bound KOR complex with $G_{i1}$ contained 220 DPPC, 60 DOPC and 120 CHL1 lipids, 43,172 water molecules, 124 sodium and 116 chloride ions. The momSalB-bound KOR complex with $G_{oA}$ contained 220 DPPC, 60 DOPC and 120 CHL1 lipids, 41,663 water molecules, and 126 sodium and 113 chloride ions. All of the systems were first processed for 50,000 steps of initial energy minimizations, then 60 ns of equilibration, followed by production runs of up to 750 ns for the KOR–$G_g$ based system and 550 ns for the rest ($G_{i1}$, $G_{oA}$ and $G_z$-bound KOR systems). The simulations were carried out on GPU clusters at the University of Southern California's High-Performance Computing Center. The temperature of 310 K and *v*-rescale thermostat algorithm were used during the production run[76]. The analyses of molecular dynamics trajectories were performed using the GROMACS software package[68].

## Data statistical analysis

For BRET2 and cAMP-inhibition assays, in the case of more than two groups, log-transformed $EC_{50}$ values were first analysed using one-way ANOVA. If significant, the Dunnett's multiple-comparison test was used to compare each mutant with the wild-type one, and the Tukey's multiple-comparison test was used to compare log-transformed $EC_{50}$ values between each group. In the case of two groups, log-transformed $EC_{50}$ values were analysed using unpaired two-tailed Student's *t*-tests to compare each mutant with a wild-type receptor. For the cell-surface expression studies, the optical density at 450 nm values of each mutant were normalized to the wild-type KOR receptor (normalized as 100%), and the resultant values were then first analysed using one-way ANOVA. If significant, a Dunnett's multiple-comparison test was used to compare each mutant with the wild-type receptor. For G-protein expression studies, the Rluc values of each mutant were normalized to the wild-type G protein (normalized as 100%), and the resultant values were then first analysed using one-way ANOVA. If significant, a Dunnett's multiple-comparison test was used to compare each mutant with the wild-type G protein. For radioligand binding and GTP turnover assays, data were analysed using unpaired two-tailed Student's *t*-

tests. In one-way ANOVA and unpaired two-tailed Student's *t*-test analysis, the significance threshold was set at $\alpha = 0.05$. Asterisks denote statistical significance; *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$; ****$P < 0.0001$; NS represents not significant.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The coordinate and cryo-EM map of KOR–$G_{i1}$–momSalB, KOR–$G_{oA}$–momSalB, KOR–$G_z$–GR89,696 and KOR–$G_g$–GR89,696 have been deposited at the PDB and Electron Microscopy Data Bank under accession codes 8DZP (EMD-27804), 8DZQ (EMD-27805), 8DZS (EMD-27807) and 8DZR (EMD-27806), respectively. All data supporting the findings of this study are available within the Article and its Supplementary Information.

## References


1. Chavkin, C. The therapeutic potential of κ-opioids for treatment of pain and addiction. *Neuropsychopharmacology* **36**, 369–370 (2011).

2. Pfeiffer, A., Brantl, V., Herz, A. & Emrich, H. M. Psychotomimesis mediated by kappa opiate receptors. *Science* **233**, 774–776 (1986).

3. Bruchas, M. R. & Roth, B. L. New technologies for elucidating opioid receptor function. *Trends Pharmacol. Sci.* **37**, 279–289 (2016).

4. Che, T. & Roth, B. L. Structural insights accelerate the discovery of opioid alternatives. *Annu. Rev. Biochem.* https://doi.org/10.1146/annurev-biochem-061620-044044 (2021).

5. Ivanina, T. et al. $G\alpha_{i1}$ and $G\alpha_{i3}$ differentially interact with, and regulate, the G protein-activated $K^+$ channel. *J. Biol. Chem.* **279**, 17260–17268

(2004).

6. Jeong, S. W. & Ikeda, S. R. G protein alpha subunit $G\alpha_z$ couples neurotransmitter receptors to ion channels in sympathetic neurons. *Neuron* **21**, 1201–1212 (1998).

7. Wong, G. T., Gannon, K. S. & Margolskee, R. F. Transduction of bitter and sweet taste by gustducin. *Nature* **381**, 796–800 (1996).

8. Wall, M. J. et al. Selective activation of Galphaob by an adenosine A1 receptor agonist elicits analgesia without cardiorespiratory depression. *Nat. Commun.* **13**, 4150 (2022).

9. Beck, T. C., Hapstack, M. A., Beck, K. R. & Dix, T. A. Therapeutic potential of kappa opioid agonists. *Pharmaceuticals* https://doi.org/10.3390/ph12020095 (2019).

10. Peet, M. M. & Baker, L. E. Salvinorin B derivatives, EOM-Sal B and MOM-Sal B, produce stimulus generalization in male Sprague–Dawley rats trained to discriminate salvinorin A. *Behav. Pharmacol.* **22**, 450–457 (2011).

11. Hayes, A. G. et al. A series of novel, highly potent and selective agonists for the κ-opioid receptor. *Br. J. Pharmacol.* **101**, 944–948 (1990).

12. Koehl, A. et al. Structure of the micro-opioid receptor–$G_i$ protein complex. *Nature* **558**, 547–552 (2018).

13. Che, T. et al. Structure of the nanobody-stabilized active state of the kappa opioid receptor. *Cell* **172**, 55–67 (2018).

14. Rasmussen, S. G. et al. Crystal structure of the $\beta_2$ adrenergic receptor–$G_s$ protein complex. *Nature* **477**, 549–555 (2011).

15. Huang, W. J. et al. Structural insights into mu-opioid receptor activation. *Nature* **524**, 315–321 (2015).

16. Rasmussen, S. G. et al. Structure of a nanobody-stabilized active state of the $\beta_2$ adrenoceptor. *Nature* **469**, 175–180 (2011).

17. Roth, B. L. et al. Salvinorin A: a potent naturally occurring nonnitrogenous kappa opioid selective agonist. *Proc. Natl Acad. Sci. USA* **99**, 11934–11939 (2002).

18. Giroud, C. et al. Salvia divinorum: an hallucinogenic mint which might become a new recreational drug in Switzerland. *Forensic Sci. Int.* **112**, 143–150 (2000).

19. Kivell, B. M., Ewald, A. W. & Prisinzano, T. E. Salvinorin A analogs and other κ-opioid receptor compounds as treatments for cocaine abuse. *Adv. Pharmacol.* **69**, 481–511 (2014).

20. Wang, Y. et al. 2-Methoxymethyl-salvinorin B is a potent kappa opioid receptor agonist with longer lasting action in vivo than salvinorin A. *J. Pharmacol. Exp. Ther.* **324**, 1073–1083 (2008).

21. Butelman, E. R. et al. GR89,696: a potent kappa-opioid agonist with subtype selectivity in rhesus monkeys. *J. Pharmacol. Exp. Ther.* **298**, 1049–1059 (2001).

22. Wang, Y. et al. Structures of the entire human opioid receptor family. *Cell* **186**, 413–427 (2023).

23. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **25**, 366–428 (1995).

24. Vardy, E. et al. A new DREADD facilitates the multiplexed chemogenetic interrogation of behavior. *Neuron* **86**, 936–946 (2015).

25. Vardy, E. et al. Chemotype-selective modes of action of κ-opioid receptor agonists. *J. Biol. Chem.* **288**, 34470–34483 (2013).

26. Wu, H. et al. Structure of the human κ-opioid receptor in complex with JDTic. *Nature* **485**, 327–332 (2012).

27. Vortherms, T. A., Mosier, P. D., Westkaemper, R. B. & Roth, B. L. Differential helical orientations among related G protein-coupled receptors provide a novel mechanism for selectivity. *J. Biol. Chem.* **282**, 3146–3156 (2007).

28. Hauser, A. S. et al. GPCR activation mechanisms across classes and macro/microscales. *Nat. Struct. Mol. Biol.* **28**, 879–888 (2021).

29. Zarzycka, B., Zaidi, S. A., Roth, B. L. & Katritch, V. Harnessing ion-binding sites for GPCR pharmacology. *Pharmacol. Rev.* **71**, 571–595 (2019).

30. Olsen, R. H. J. et al. TRUPATH, an open-source biosensor platform for interrogating the GPCR transducerome. *Nat. Chem. Biol.* **16**, 841–849 (2020).

31. Mueller, K. L. et al. The receptors and coding logic for bitter taste. *Nature* **434**, 225–229 (2005).

32. Che, T. et al. Nanobody-enabled monitoring of kappa opioid receptor states. *Nat. Commun.* **11**, 1145 (2020).

33. Flock, T. et al. Universal allosteric mechanism for Galpha activation by GPCRs. *Nature* **524**, 173–179 (2015).

34. Mafi, A., Kim, S. K. & Goddard III, W. A. The mechanism for ligand activation of the GPCR-G protein complex. *Proc. Natl Acad. Sci. USA* **119**, e2110085119 (2022).

35. Rose, A. S. et al. Position of transmembrane helix 6 determines receptor G protein coupling specificity. *J. Am. Chem. Soc.* **136**, 11244–11247 (2014).

36. Glukhova, A. et al. Rules of engagement: GPCRs and G proteins. *ACS Pharmacol. Transl. Sci.* **1**, 73–83 (2018).

37. Flock, T. et al. Selectivity determinants of GPCR-G-protein binding. *Nature* **545**, 317–322 (2017).

38. Inoue, A. et al. Illuminating G-protein-coupling selectivity of GPCRs. *Cell* **177**, 1933–1947 (2019).

39. Garcia-Nafria, J., Nehme, R., Edwards, P. C. & Tate, C. G. Cryo-EM structure of the serotonin 5-HT1B receptor coupled to heterotrimeric $G_o$. *Nature* **558**, 620–623 (2018).

40. De Lean, A., Stadel, J. M. & Lefkowitz, R. J. A ternary complex model explains the agonist-specific binding properties of the adenylate cyclase-coupled beta-adrenergic receptor. *J. Biol. Chem.* **255**, 7108–7117 (1980).

41. Samama, P., Cotecchia, S., Costa, T. & Lefkowitz, R. J. A mutation-induced activated state of the $\beta_2$-adrenergic receptor. Extending the ternary complex model. *J. Biol. Chem.* **268**, 4625–4636 (1993).

42. Staus, D. P. et al. Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation. *Nature* **535**, 448–452 (2016).

43. Christopoulos, A. & Kenakin, T. G protein-coupled receptor allosterism and complexing. *Pharmacol. Rev.* **54**, 323–374 (2002).

44. Dror, R. O. et al. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* **348**, 1361–1365 (2015).

45. Gregorio, G. G. et al. Single-molecule analysis of ligand efficacy in $\beta_2$AR–G-protein activation. *Nature* **547**, 68–73 (2017).

46. Chung, K. Y. et al. Conformational changes in the G protein $G_s$ induced by the $\beta_2$ adrenergic receptor. *Nature* **477**, 611–615 (2011).

47. Liang, Y. L. et al. Dominant negative G proteins enhance formation and purification of agonist-GPCR-G protein complexes for structure determination. *ACS Pharmacol. Transl. Sci.* **1**, 12–20 (2018).

48. Gilman, A. G. G proteins: transducers of receptor-generated signals. *Annu. Rev. Biochem.* **56**, 615–649 (1987).

49. Seyedabadi, M., Gharghabi, M., Gurevich, E. V. & Gurevich, V. V. Structural basis of GPCR coupling to distinct signal transducers: implications for biased signaling. *Trends Biochem. Sci.* https://doi.org/10.1016/j.tibs.2022.03.009 (2022).

50. Cong, X. et al. Molecular insights into the biased signaling mechanism of the mu-opioid receptor. *Mol. Cell* **81**, 4165–4175 (2021).

51. Suomivuori, C. M. et al. Molecular mechanism of biased signaling in a prototypical G protein-coupled receptor. *Science* **367**, 881–887 (2020).

52. Lane, J. R., May, L. T., Parton, R. G., Sexton, P. M. & Christopoulos, A. A kinetic view of GPCR allostery and biased agonism. *Nat. Chem. Biol.* **13**, 929–937 (2017).

53. Grundmann, M. & Kostenis, E. Temporal bias: time-encoded dynamic GPCR signaling. *Trends Pharmacol. Sci.* **38**, 1110–1124 (2017).

54. Wenzel-Seifert, K. & Seifert, R. Molecular analysis of $\beta_2$-adrenoceptor coupling to $G_s$-, $G_i$-, and $G_q$-proteins. *Mol. Pharmacol.* **58**, 954–966 (2000).

55. Wall, M. A. et al. The structure of the G protein heterotrimer $G_{i\alpha 1}\beta_1\gamma_2$. *Cell* **83**, 1047–1058 (1995).

56. Maeda, S. et al. Development of an antibody fragment that stabilizes GPCR/G-protein complexes. *Nat. Commun.* **9**, 3712 (2018).

57. Lee, D. Y. et al. Synthesis and in vitro pharmacological studies of new C(2) modified salvinorin A analogues. *Bioorg. Med. Chem. Lett.* **15**, 3744–3747 (2005).

58. Munro, T. A. et al. Standard protecting groups create potent and selective kappa opioids: salvinorin B alkoxymethyl ethers. *Bioorg. Med. Chem.* **16**, 1279–1286 (2008).

59. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51

(2005).

60. Peck, J. V., Fay, J. F. & Strauss, J. D. High-speed high-resolution data collection on a 200 keV cryo-TEM. *IUCrJ* **9**, 243–252 (2022).

61. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

62. Punjani, A., Zhang, H. & Fleet, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nat. Methods* **17**, 1214–1221 (2020).

63. Cao, C. et al. Structure, function and pharmacology of human itch GPCRs. *Nature* **600**, 170–175 (2021).

64. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

65. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).

66. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).

67. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).

68. Abraham, M. J. et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).

69. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).

70. Kim, S. et al. CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules. *J. Comput. Chem.* **38**, 1879–1886 (2017).

71. Abagyan, R., Totrov, M. & Kuznetsov, D. ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **15**, 488–506 (1994).

72. Xing, C. et al. Cryo-EM structure of the human cannabinoid receptor CB2–$G_i$ signaling complex. *Cell* **180**, 645–654 (2020).

73. Arnautova, Y. A., Abagyan, R. A. & Totrov, M. Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling. *Proteins Struct. Funct. Bioinform.* **79**, 477–498 (2011).

74. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370–D376 (2012).

75. Leonard, A. N. & Lyman, E. Activation of G-protein-coupled receptors is thermodynamically linked to lipid solvation. *Biophys. J.* **120**, 1777–1787 (2021).

76. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* https://doi.org/10.1063/1.2408420 (2007).

77. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

## Acknowledgements

We thank J. Peck and J. Strauss for technical assistance in this project; the staff at the Washington University Center for Cellular Imaging for sample screening; the staff of the NIDA Drug Supply Program for providing the $^3$H-JDTic; X.-P. Huang and the members of the Psychoactive Drug Screening

Program (PDSP) at UNC-Chapel Hill for the GPCRome screening analysis; and staff at the Center for Advanced Research Computing (CARC) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication. The work is supported by NIH grants R35GM143061 (to T.C.) and R01NS099341 (to P.Y.). The Titan X Pascal used for this research was donated to J.F.F. by NVIDIA.

# Author information

Author notes

1. Jingying Zhang & Peng Yuan

   Present address: Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

2. Jingying Zhang & Peng Yuan

   Present address: Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA

## Authors and Affiliations

1. Department of Anesthesiology, Washington University in St Louis, St Louis, MO, USA

   Jianming Han, Sarah M. Bernhard, Lei Zhao, Susruta Majumdar & Tao Che

2. Center for Clinical Pharmacology, University of Health Sciences and Pharmacy in St Louis and Washington University School of Medicine, St Louis, MO, USA

   Jianming Han, Sarah M. Bernhard, Vipin A. Rangari, Susruta Majumdar & Tao Che

3. Department of Cell Biology and Physiology, Washington University School of Medicine, St Louis, MO, USA

Jingying Zhang & Peng Yuan

4. Center for the Investigation of Membrane Excitability Diseases, Washington University School of Medicine, St Louis, MO, USA

Jingying Zhang & Peng Yuan

5. Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA

Antonina L. Nazarova, Jordy Homing Lam & Vsevolod Katritch

6. Department of Chemistry, University of Southern California, Los Angeles, CA, USA

Antonina L. Nazarova, Jordy Homing Lam & Vsevolod Katritch

7. Center for New Technologies in Drug Discovery and Development, Bridge Institute, Michelson Center for Convergent Biosciences, University of Southern California, Los Angeles, CA, USA

Antonina L. Nazarova, Jordy Homing Lam & Vsevolod Katritch

8. Department of Pharmacology, University of North Carolina School of Medicine, Chapel Hill, NC, USA

Brian E. Krumm

9. Washington University Pain Center, Washington University in St Louis, St Louis, MO, USA

Susruta Majumdar & Tao Che

10. Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA

David E. Nichols

11. Department of Biochemistry and Molecular Biology, University of Maryland Baltimore, Baltimore, MD, USA

Jonathan F. Fay

### Contributions

J.H. prepared the four KOR–G protein complexes, performed the functional validation assays and prepared the manuscript. J.Z. helped with the cryo-EM data processing. A.L.N. and J.H.L. performed the molecular dynamics simulations. S.M.B. helped with construct optimization, sample preparation and saturation binding experiments. B.E.K. provided the expression plasmid for $G_g$. L.Z. helped with protein expression. V.A.R. and S.M. helped to collect the NMR spectrum of momSalB. D.E.N. synthesized the compound momSalB. V.K. supervised the molecular dynamics simulations and manuscript discussion. P.Y. helped with the map discussion. J.F.F. prepared grids, collected the cryo-EM data and resolved the cryo-EM density map. P.Y., J.F.F. and T.C. prepared the manuscript and supervised the project.

### Corresponding authors

Correspondence to Jonathan F. Fay or Tao Che.

## Ethics declarations

### Competing interests

The authors declare no competing interests.

## Peer review

### Peer review information

*Nature* thanks Graham Ladds and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

## Additional information

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Extended data figures and tables

### Extended Data Fig. 1 Comparison of KOR-Gi1 complex structure with MOR-Gi1 and active-state KOR-Nb39 structures.

**a**. Opioid receptors primarily couple to Gi/o family subtypes upon activation. **b**. Sequence alignment (αN and α5 helices) of Gαi/o family subtypes. The percentage represents the sequence similarity related to Gi1 (set as 100%). **c**. Overall alignment of KOR-Gi1 and MOR-Gi1 structures. The two structures are globally similar to each other, including TM6 of the receptor and α5 helix of the Gα subunit, but different in the αN helix of Gα. **d**. The Gi1-bound KOR differs from the Nb39-bound KOR in the degrees of TM6 outward movement (by 2 Å). **e**. Both Gi1 and Nb39 act as positive allosteric modulators of KOR, but differentially increase the binding affinity of KOR agonists. Data are grouped data ± s.e.m. from n = 3 biological replicates. Full quantitative parameters from this experiment are listed in Supplementary Table 8.

### Extended Data Fig. 2 The binding pharmacology of momSalB and GR89,696.

**a**. GPCRome screening at 320 GPCRs that measures agonist activity of tested ligands shows that momSalB and GR89,696 are selective at KOR. **b**. The binding poses of momSalB and GR89,696 show that they adopt different planes in the orthosteric pocket. **c**. A cartoon model of G protein-mediated cAMP reporter assays. The Gαi/o here represents all subtypes

expressed in the cells. **d**. Mutagenesis screening of key binding-pocket residues using G protein-mediated cAMP reporter assay. Data are grouped data ± s.e.m. of n = 3 biological replicates. Full quantitative parameters from this experiment are listed in Supplementary Table 2. **e**. Measurement of cell surface expression of KOR binding pocket mutants by ELISA. In general, these mutants maintained robust cell surface expression. Although some mutations significantly altered surface expression compared to the wild type, the increased or decreased expression appeared to minimally affect the potency in agonist-mediated cAMP inhibition. Bar-graphs are $OD_{450}$ ± s.e.m. from n = 3 biological replicates. Statistical significance for each mutant is compared in a one-way analysis of variance (ANOVA) with Dunnett's multiple comparisons test to the wild type (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$, "ns" represents no significance; V108A: $p = 0.3919$, W124A: $p = 0.0893$, V134A: $p = 0.0308$, M142A: $p = 0.0398$, V230A: $p = 0.0004$, H291A: $p = 0.0002$, Y320L: $p = 0.0008$). Signalling curves are grouped data ± s.e.m. of n = 3 biological replicates. Full quantitative parameters from this experiment are listed in Supplementary Table 9.

## Extended Data Fig. 3 Molecular determinants of momSalB agonism.

**a**. The positive effect of additional D138$^{3.32}$N mutation on the momSalB-mediated cAMP inhibition through KOR. The additional D138$^{3.32}$N mutation does not rescue U50,488 or GR89,696-mediated cAMP inhibition. Data are grouped data ± s.e.m. from n = 3 biological replicates. Full quantitative parameters from this experiment are listed in Supplementary Table 10. **b**. Effects of mutations in the hydrophobic pocket on the binding affinity of momSalB. Data are grouped data ± s.e.m. from n = 3 independent biological replicates. Full quantitative parameters from this experiment are listed in Supplementary Table 11. **c** and **d**. Chemical structures of SalB, momSalB, and EOM-SalB. Differences are highlighted by red colour. The agonist activity of each analogue is shown in the parentheses. Data for EOM-SalB was taken from ref. [19]. Binding poses of SalB and EOM-SalB at KOR were revealed by molecular docking performed in the Schrodinger Maestro v12.9. The three ligands occupy a similar binding pocket with different extents toward the hydrophobic pocket.

## Extended Data Fig. 4 Comparison of interface details and areas between KOR and each G protein subtypes.

Specific interactions between the receptor and individual Gα subunits, Gi1 (**a**), GoA (**b**), Gz (**c**) and Gg (**d**). (Cartoon) Key residues from the intracellular side of KOR, and residues in the αN and α5 helix were mapped out. The H-bond or salt-bridge interactions are shown as red dashed lines. The closest distances between the intracellular KOR residues and the Gα residues were labelled. The distance cutoff is 4 Å. The interface area was calculated by the online server PDBePISA[77].

## Extended data Fig. 5 The effect of KOR-G-protein interface residues on the G protein coupling.

**a**. Mutagenesis screening of KOR-G protein interface (KOR side) residues by cAMP inhibition assays. Data are grouped data ± s.e.m. from n = 3 biological replicates. **b**. Mutagenesis analysis of intracellular KOR residues by cAMP inhibition assays. Data are mean LogEC50 ± s.e.m. from n = 3 biological replicates. Statistical significance for each mutant is compared in a one-way analysis of variance (ANOVA) with Dunnett's multiple comparisons test to the wild type (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$, "ns" represents no significance, GR89,696: V160A: $p = 0.0001$, P163A: $p = 0.0020$, R252A: $p = 0.0014$, L253A: $p = 0.0003$, R257A: $p = 0.0003$, R271A: $p = 0.0004$; momSalB: L167A: $p = 0.0003$, R257A: $p = 0.1514$, R271A: $p = 0.0004$, I272A: $p = 0.0002$). Full quantitative parameters from this experiment are listed in Supplementary Table 12. **c**. Measurement of cell surface expression of KOR-G protein interface mutants by ELISA. The BRET results suggest a minor effect with different concentrations of KOR plasmids. Bar-graphs are $OD_{450}$ ± s.e.m. from n = 3 independent biological replicates. Statistical significance for each mutant is compared in a one-way analysis of variance (ANOVA) with the Dunnett's multiple comparisons test to the wild type (* = $p < 0.05$, ** = $p<0.01$, *** = $p < 0.001$, **** = $p < 0.0001$, "ns" represents no significance; V160A: $p = 0.2732$, P163A: $p = 0.9992$, R252A: $p = 0.9998$, L253A: $p = 0.8943$, R257A: $p = 0.9997$, N336A: $p = 0.9992$). Signalling curves are grouped data ± s.e.m. of n = 3 biological replicates. Full quantitative parameters from this experiment are listed in Supplementary Table 13. **d**.

The KOR-R156A mediated decrease of efficacy was confirmed by the BRET2-based kinetic measurement. Data are net BRET ± s.e.m. from n = 3 independent biological replicates.

## Extended Data Fig. 6 Molecular dynamics (MD) simulations reveal distance traces of interactions between the KOR residues (R156$^{3.50}$) with Gα protein residues.

MD simulations revealed the distance traces and frequency between KOR-R156$^{3.50}$ and L353/354$^{H5.25}$ in the α5 of each Gα subunit. The position of KOR-R156$^{3.50}$ is also supported by the stable interaction with KOR-Y246$^{5.58}$. Five independent simulations of KOR-G complex (coral, orange, green, cyan, blue) are shown, spanning 0.55-0.75 µs of cumulative time per system, with the sampling rate of 10 frames per ns, solid lines and same-colour shadows representing moving average values and one standard deviation respectively from 50 frames in all cases.

## Extended Data Fig. 7 Molecular dynamics (MD) simulations reveal distance traces of interactions between the KOR residues (N336$^{8.49}$) with Gα protein residues.

The closest distances of polar residues N336$^{8.49}$ in KOR with residues in Gi1, GoA, Gz, or Gg are shown. Distance histograms for each plot are also shown in parallel. Five independent simulations of KOR-G complex (coral, orange, green, cyan, blue) are shown, spanning 0.55 µs (KOR-Gi1, GoA, Gz) and 0.75 µs (KOR-Gg) of cumulative time per system, with the sampling rate of 10 frames per ns, solid lines and same-colour shadows representing moving average values and one standard deviation respectively from 50 frames in all cases.

## Extended Data Fig. 8 The role of G protein interface residues on KOR-G protein coupling.

**a**. The nonconserved residues in the G protein α5 and αN helices. **b-e**. The effects of individual mutations in each G protein subtype were screened

using the BRET2 KOR-G protein assays. Data are grouped data ± s.e.m. from n = 3 biological replicates. Full quantitative parameters from this experiment are listed in Supplementary Table 14. **f**. The expression levels of G protein mutants were quantified based on the luminescence counts from the Gα-Rluc excitation. Data are BRET ratio ± s.e.m. from n = 3 biological replicates. Statistical significance for each mutant is compared in a one-way analysis of variance (ANOVA) with Dunnett's multiple comparisons test to the wild type (* = $p < 0.05$, ** = $p<0.01$, *** = $p < 0.001$, **** = $p < 0.0001$, "ns" represents no significance; Gi1-K345A: $p = 0.0015$, Gi1-C351A: $p = 0.7984$; GoA-A345Q: $p = 0.0210$, GoA-R349A: $p = 0.0019$, GoA-G350A: $p = 0.0284$, GoA-C351A: $p = 0.1102$, GoA-Y354A: $p = 0.6955$; Gz-R31A: $p > 0.9999$, Gz-Y351A: $p = 0.9947$, Gz-I352A: $p = 0.5286$, Gz-C355A: $p = 0.0161$; Gg-K345A: $p = 0.5849$, Gg-E346A: $p = 0.0050$, Gg-D350A: $p = 0.9936$, Gg-C351A: $p = 0.9808$, Gg-F354A: $p = 0.0091$).

## Extended Data Fig. 9 Comparison of KOR-coupled Gα subunit with other G protein families.

**a**. The KOR-GoA displays a different conformation from the 5-HT1BR-miniGo structure, including both TM6 of the receptor and α5 helix of the Gα subunit. **b**. Comparison of receptor-G protein interface among Gs, Gi1, Gq bound complexes. The TM6 of the receptor, αN and α5 helices of the Gα subunits adopt different conformations. **c**. The receptor-Gα interface of β2AR-Gαs and 5-HT2AR-Gαq. The dashed lines represent the closest distance between the intracellular receptor residues and the Gα residues. The distance cutoff is 4 Å. The interface area shown in the brackets was calculated by the online server PDBePISA.

## Extended Data Fig. 10 The presence of GDP or GTP reduces the allosteric potentiation of GoA, Gz, or Gg, respectively.

**a**. Comparison of functional activity between wild-type G proteins and engineered G proteins by GTP turnover assay. Data are luminescence ± s.e.m. of n = 3 biological replicates with each performed in triplicate. Significance analyses were performed using unpaired two-tailed student's t-test to compare each point in engineered G protein (G protein-DN) to the corresponding point in wild-type G proteins. (* = $p < 0.05$, ** = $p < 0.01$,

*** = $p < 0.001$, **** = $p < 0.0001$). **b**. Measurement of allosteric potentiation of engineered G proteins on KOR agonist binding. The concentration of $^{3}$H-U69,593 was 8 nM, and the G protein was 250 nM. DN, dominant negative. Data are cpm ± s.e.m. from n = 4 biological replicates. Significance analyses were performed using the unpaired two-tailed student's t-test to compare each group to the KOR-only group (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$, "ns" represents no significance; (KOR+GoAwt) vs KOR: $p = 0.0002$, (KOR+Ggwt) vs KOR: $p = 0.2469$, (KOR+Gi1DN) vs KOR: $p = 0.0012$, (KOR+GoADN) vs KOR: $p = 0.0007$, (KOR+GzDN) vs KOR: $p = 0.0029$, (KOR+GgDN) vs KOR: $p = 0.0719$). **c**. The saturation binding assays were conducted with or without 50 μM GDP/GTP. Data are Bmax ± s.e.m. from n = 3 biological replicates with each performed in duplicate. Statistical significance analyses between groups (with GTP/GDP and without GTP/GDP) are compared in the unpaired two-tailed student's t-test (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$; (KOR+GoA) vs (KOR+GoA+GTP): $p = 0.0001$, (KOR+Gg) vs (KOR+Gg+GDP): $p = 0.0008$, (KOR+Gg) vs (KOR+Gg+GTP): $p = 0.0010$). Full quantitative parameters from this experiment are listed in Supplementary Table 15.

### Extended Data Fig. 11 Comparison of uncoupled GDP-bound Gi1 and KOR-coupled nucleotide-free Gi1.

**a**. Comparison of GDP-bound Gi1 with KOR bound Gi1, GoA, Gz, and Gg. **b**. Conformational translocation of α1 and α5 helices upon GPCR engagement. **c**. α5 helix shows an 8.6 Å upward movement into the KOR intracellular core.

**Extended Data Table 1 Cryo-EM data collection, refinement, and validation statistics**

## Supplementary information

### Supplementary Information

Supplementary Methods (allosteric $IC_{50}$ shift model equation), Supplementary Figs. 1–6 and Supplementary Tables 1–17.

**Reporting Summary**

**Peer Review File**

## Rights and permissions

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Reprints and Permissions

## About this article

### Cite this article

Han, J., Zhang, J., Nazarova, A.L. *et al.* Ligand and G-protein selectivity in the κ-opioid receptor. *Nature* **617**, 417–425 (2023). https://doi.org/10.1038/s41586-023-06030-7

- Received: 14 August 2022
- Accepted: 29 March 2023

- Published: 03 May 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-06030-7

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-06030-7

# Amendments & Corrections

- **Author Correction: Basin-scale reconstruction of euxinia and Late Devonian mass extinctions** [05 Apr 2023]
  - 
- **Addendum: Agrochemicals interact synergistically to increase bee mortality** [24 Apr 2023]
  - 

- Author Correction
- Published: 05 April 2023

# Author Correction: Basin-scale reconstruction of euxinia and Late Devonian mass extinctions


- Swapan K. Sahoo ,
- Geoffrey J. Gilleaudeau,
- Kathleen Wilson,
- Bruce Hart,
- Ben D. Barnes,
- Tytrice Faison,
- Andrew R. Bowman,
- Toti E. Larson &
- …
- Alan J. Kaufman

Show authors


*Nature* **volume 617**, page E6 (2023)


- 460 Accesses

- 1 Altmetric

- Metrics details


The Original Article was published on 08 March 2023

Correction to: *Nature* https://doi.org/10.1038/s41586-023-05716-2
Published online 8 March 2023

In the version of this article originally published, the surname of Toti E. Larson was misspelled as Larsen and has been corrected in the HTML and PDF versions of the article.

# Author information

Author notes

1. These authors contributed equally: Swapan K. Sahoo, Geoffrey J. Gilleaudeau, Alan J. Kaufman

## Authors and Affiliations

1. Equinor US, Houston, TX, USA

   Swapan K. Sahoo & Bruce Hart

2. George Mason University, Fairfax, VA, USA

   Geoffrey J. Gilleaudeau

3. Boston College, Chestnut Hill, MA, USA

   Kathleen Wilson

4. University of Western Ontario, London, Ontario, Canada

   Bruce Hart

5. Pennsylvania State University, University Park, PA, USA

   Ben D. Barnes

6. University of Maryland, College Park, MD, USA

Tytrice Faison & Alan J. Kaufman

7. Gaia Paleontological Services LLC, Missouri City, TX, USA

Andrew R. Bowman

8. Bureau of Economic Geology, University of Texas at Austin, Austin, TX, USA

Toti E. Larson

### Corresponding author

Correspondence to Swapan K. Sahoo.

## Rights and permissions

Reprints and Permissions

## About this article

### Cite this article

Sahoo, S.K., Gilleaudeau, G.J., Wilson, K. *et al.* Author Correction: Basin-scale reconstruction of euxinia and Late Devonian mass extinctions. *Nature* **617**, E6 (2023). https://doi.org/10.1038/s41586-023-05999-5

- Published: 05 April 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05999-5

### Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05999-5

| Section menu | Main menu |

- Addendum
- Published: 24 April 2023

# Addendum: Agrochemicals interact synergistically to increase bee mortality


- Harry Siviter ,
- Emily J. Bailes,
- Callum D. Martin,
- Thomas R. Oliver,
- Julia Koricheva,
- Ellouise Leadbeater &
- …
- Mark J. F. Brown

Show authors


*Nature* **volume 617**, pages E7–E9 (2023)


- 512 Accesses

- 2 Altmetric

- Metrics details

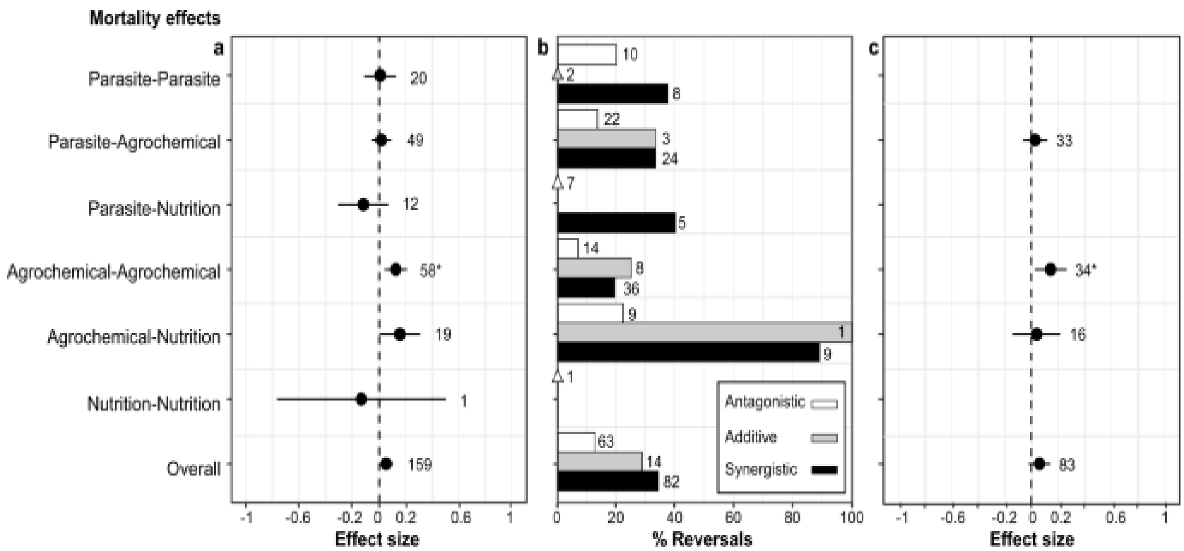

The Original Article was published on 04 August 2021

Addendum to: *Nature* https://doi.org/10.1038/s41586-021-03787-7
Published online 4 August 2021

In our original article, we assessed whether interactions between stressors that impact bee health were antagonistic, additive or synergistic. To do this, we followed the methods developed by Jackson et al.[1] and determined the overall interaction effect as the standardized mean difference (SMD) (Hedges' *d*) between the predicted additive effect and observed effect of both stressors. Since then, it has been brought to our attention that (i) there were errors in our uploaded dataset, (ii) the formula for the sampling variance of the interaction effects in Gurevitch et al.[2] was more appropriate for our dataset than the one we had previously used (see ref. [1]) (with thanks to Y. Yefeng and S. Nagakawa for this personal communication), and (iii) it is possible to use proportion difference approaches for this kind of interaction analysis, rather than the SMD approach we used. We have corrected our analyses with respect to these points, and the details of this are outlined below.

Dataset: an error in the dataset initially uploaded disassociated a subset of means from their standard deviations. We have corrected this and, in the process, reassessed each row of data against the original paper, removing or updating any instances where errors had been made in either data extraction or study interpretation ($n = 75$). This resulted in the $n$ values of each analysis changing from the values seen in the original analysis (Fig. 1). The correct datasets have now been uploaded to the original repository. However, this final dataset is not our complete dataset, as 63 unpublished effect sizes were confidentially shared with us by the authors of these studies.

**Fig. 1: Revised — The interaction effects of parasites, agrochemicals, and nutritional stressors on bee mortality.**




**a**, The effect sizes for the indicated interaction effects. Data are PD values ± 95% CI. Interactions are synergistic when the effect size is positive and the 95% CI does not include zero, antagonistic when the effect size is negative and the 95% CI does not include zero, and additive when the 95% CI includes zero. **b**, The percentage of additive, antagonistic and synergistic interactions between stressors that were reversal interactions (Methods). **c**, Effect sizes (PD values ± 95% CI) for analyses in which bees are exposed to field-realistic concentrations of agrochemicals.

Sampling variance calculation: we recalculated, where appropriate (see below), the variance for interaction effects using the formula reported in Gurevitch et al.[2].

SMD: mortality data are proportion data and, as such, the appropriate meta-analytical approach is to use the proportion difference rather than the SMD[3]. We originally used SMD in the absence of a method for calculating variance that could be applied to interaction effects derived from proportion data[1]. In this correction, we now use proportion difference with an appropriate variance calculation (with thanks to S. Nagakawa for deriving this). In addition, because our dataset is a mixture of studies where standard deviations could be extracted at the cage level or the individual level, we calculated the standard deviation differently for each type of data. For cage-level data, we used the extracted standard deviations from the paper to calculate the sampling variance, as reported in our original manuscript. For individual-level data, we calculated standard deviation (*s*) using the formula

$$s=\sqrt{\frac{{pq}}{n}}$$

where $p$ is the number of surviving individuals, $q$ is $(1 - p)$ and $n$ is the total number of individuals.

Finally, to improve the robustness of our analysis, in the updated analysis we (i) included ‘stressor type’ as a moderator in the main meta-analytic model as opposed to sub-setting the data, and (ii) used cage-level data wherever possible (e.g., the data provided in the Supplementary Material 19 data points were re-extracted at cage level). The results of analyses using the above approach, but with data extracted at the individual level when available (as reported in the paper initially published), were qualitatively the same as our new analyses when cage-level data were used. The only exception was that the overall interaction effect remained synergistic, not additive (PD, proportion difference = 0.05; CI, 95% confidence intervals = 0.001 to 0.09 — see below for comparison with the analysis incorporating available cage-level data).

We report the reanalyses of our published results below.

**Agrochemicals interact synergistically to increase bee mortality**. While the overall interaction between stressors was additive in the revised analyses (PD = 0.05, CI = −0.001 to 0.09, $n = 159$), agrochemicals still showed a synergistic interaction (PD = 0.12, CI = 0.04 to 0.20, $n = 58$) (Fig. 1a). Thus, our key finding that agrochemicals interact synergistically to increase bee mortality does not change. All other interactions remained additive (parasite × parasite, PD = 0.00, CI = −0.12 to 0.12, $n = 20$; parasite × agrochemical, PD = 0.02, CI = −0.06 to 0.08, $n = 49$; parasite × nutrition, PD = −0.12, CI −0.32 to 0.06, $n = 12$; agrochemical × nutrition, PD = 0.15, CI −0.00 to 0.30, $n = 19$; nutrition × nutrition, PD = −0.14, CI = −0.78 to 0.50, $n = 1$). This pattern was mirrored in analyses where only studies using field-realistic agrochemical exposure were used (Fig. 1c).

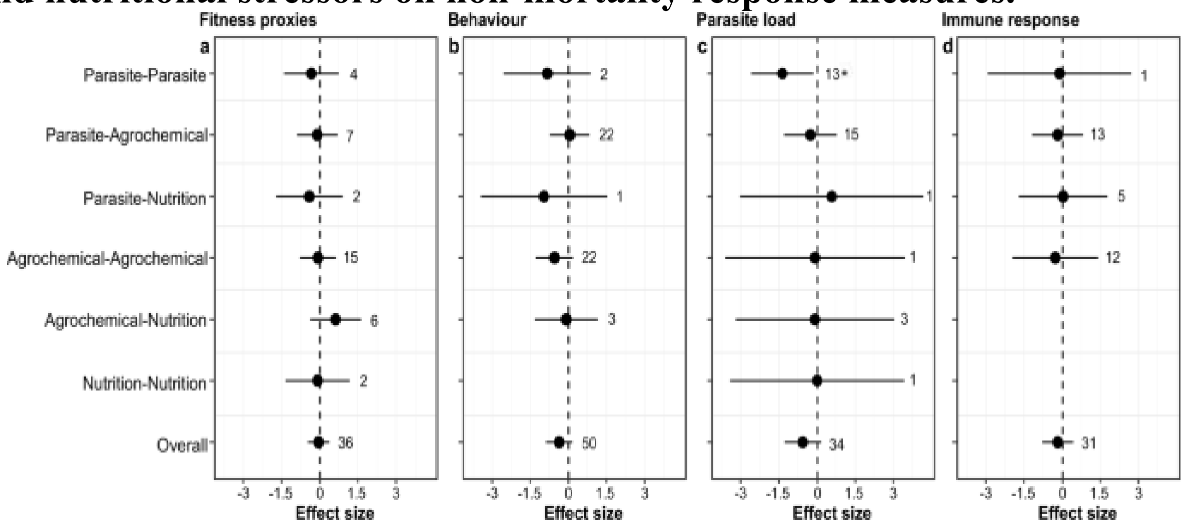
We also found the same synergistic interactions between specific agrochemicals, as observed in the original analysis (Extended Data Fig. 1, which is a revised version of the original Extended Data Fig. 4).

**Interaction effects on non-mortality end-points are largely additive**. We reanalysed our non-mortality datasets using the Gurevitch et al.[2] formula for sampling variance, pooled standard deviation and *j*:

$$j=1-(3/(4\times ({\rm{crossed}}\,n+{\rm{stress}}\,1\,n+{\rm{stress}}\,2\,n+{\rm{control}}\,n-4)-1)).$$

To avoid mixing dependent variables with different error distributions, only studies with Gaussian data distributions were included ($n = 28$ effect sizes removed). Again, the results were largely qualitatively similar (Fig. 2).

**Fig. 2: Revised — The interaction effects of parasites, agrochemicals and nutritional stressors on non-mortality response measures.**




**a**–**d**, The interaction effects of parasites, agrochemicals and nutritional stressors on (**a**) bee fitness proxies, (**b**) bee behaviour, (**c**) parasite load and (**d**) immune response. The effect sizes for the indicated interaction effects. Data are Hedges' *d* values ± 95% CI. Interactions are synergistic when the effect size is positive and the 95% CI does not include zero, antagonistic when the effect size is negative and the 95% CI does not include zero, and additive when the 95% CI includes zero.

Fitness: the results mirrored those of the original analyses. The overall interaction between stressors was additive ($d = -0.04$, CI = −0.46 to 0.39, $n = 36$), as were the individual interactions (parasite × parasite, $d = -0.32$, CI = −1.40 to 0.76, $n = 4$; parasite × agrochemical, $d = -0.09$, CI = −0.88 to 0.70, $n = 7$; parasite × nutrition, $d = -0.40$, CI = −1.71 to 0.91, $n = 2$;

agrochemical × agrochemical, $d = -0.09$, CI = −0.88 to 0.70, $n = 15$; agrochemical × nutrition, $d = 0.63$, CI = −0.36 to 1.62, $n = 6$; nutrition × nutrition, $d = -0.08$, CI = −1.33 to 1.18, $n = 2$).

Behaviour: the overall interaction between stressors was additive ($d = -0.36$, CI = −0.88 to 0.17, $n = 50$), as opposed to the significant antagonistic interaction initially reported in the paper. However, individual interactions largely remained additive as in the original analysis (parasite × parasite, $d = -0.83$, CI = −2.54 to 0.89, $n = 2$; parasite × agrochemical, $d = 0.07$, CI = −0.71 to 0.85, $n = 22$; parasite × nutrition, $d = -0.96$, CI = −3.44 to 1.53, $n = 1$; agrochemical × agrochemical, $d = -0.54$, CI = −1.28 to 0.19, $n = 22$) with one interaction becoming additive, as opposed to antagonistic (agrochemical × nutrition, $d = -0.08$, CI = −1.32 to 1.16, $n = 3$).

Parasite load: in contrast to the original analysis published, the overall interaction between stressors was additive ($d = -0.56$, CI = −1.28 to 0.16, $n = 34$), rather than antagonistic. However, the parasite × parasite interaction remained antagonistic (parasite × parasite, $d = -1.36$, CI = −2.58 to −0.14, $n = 13$) and the remaining individual interactions remained additive (parasite × agrochemical, $d = -0.26$, CI = −1.30 to 0.78, $n = 15$; parasite × nutrition, $d = 0.59$, CI = −3.01 to 4.19, $n = 1$; agrochemical × agrochemical, $d = -0.08$, CI = −3.60 to 3.45, $n = 1$; agrochemical × nutrition, $d = -0.08$, CI = −3.19 to 3.04, $n = 3$; nutrition × nutrition, $d = 0.01$, CI = −3.41 to 3.43, $n = 1$).

Immune: the results of reanalyses mirrored the original analyses. The overall interaction between stressors ($d = -0.18$, CI = −0.79 to 0.43, $n = 31$) and the individual interactions were additive (parasite × parasite, $d = -0.11$, CI = −2.92 to 2.71, $n = 1$; agrochemical × parasite, $d = -0.18$, CI = −1.17 to 0.82, $n = 13$; parasite × nutrition, $d = 0.04$, CI = −1.70 to 1.78, $n = 5$; agrochemical × agrochemical, $d = -0.27$, CI = −1.94 to 1.40, $n = 12$).

Overall, despite the changes in dataset and in analytical methodology, our results and conclusions remain largely the same, particularly our key result regarding synergistic agrochemical interactions, and thus the original conclusions of the paper are upheld. We are extremely grateful to S. Nakagawa and Y. Yefeng for their help and support in correcting this paper.

# References

1. Jackson, M. C., Loewen, C. J. G., Vinebrooke, R. D. & Chimimba, C. T. Net effects of multiple stressors in freshwater ecosystems: a meta-analysis. *Glob. Change Biol.* **22**, 180–189 (2016).

2. Gurevitch, J., Morrison, J. A. & Hedges, L. V. The interaction between competition and predation: a meta-analysis of field experiments. *Am. Nat.* **155**, 435–453 (2000).

3. Rosenberg, M. S., Rothstein, H. R. & Gurevitch, J. in *Handbook of Meta-Analysis in Ecology and Evolution* (eds Koricheva, J. et al.) Ch. 6 (Princeton University Press, 2013).

# Author information

Author notes

1. These authors contributed equally: Harry Siviter, Emily J. Bailes

## Authors and Affiliations

1. Department of Biological Sciences, Royal Holloway University of London, Egham, UK

   Harry Siviter, Emily J. Bailes, Callum D. Martin, Thomas R. Oliver, Julia Koricheva, Ellouise Leadbeater & Mark J. F. Brown

2. Department of Integrative Biology, University of Texas at Austin, Austin, TX, USA

   Harry Siviter

3. Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK

   Emily J. Bailes

4. School of Natural Sciences, Bangor University, Bangor, UK

Emily J. Bailes & Thomas R. Oliver

5. Rothamsted Research, Harpenden, UK

Thomas R. Oliver

### Corresponding author

Correspondence to Harry Siviter.

## Additional information

### Extended data

is available for this paper at https://doi.org/10.1038/s41586-023-05997-7.

## Extended data figures and tables

### Extended Data Fig. 1 The interaction effects of different agrochemical classes on bee mortality response measures.

This is a revised version of the original Extended Data Fig. 4. PD ± 95% CI are shown. Note that effect sizes for azole fungicide × pyrethroid are included in both groups.

## Rights and permissions

Reprints and Permissions

## About this article

### Cite this article

Siviter, H., Bailes, E.J., Martin, C.D. *et al.* Addendum: Agrochemicals interact synergistically to increase bee mortality. *Nature* **617**, E7–E9 (2023). https://doi.org/10.1038/s41586-023-05997-7

- Published: 24 April 2023
- Issue Date: 11 May 2023
- DOI: https://doi.org/10.1038/s41586-023-05997-7

## Share this article

Anyone you share the following link with will be able to read this content:

Get shareable link

Sorry, a shareable link is not currently available for this article.

Copy to clipboard

Provided by the Springer Nature SharedIt content-sharing initiative

---

This article was downloaded by **calibre** from https://www.nature.com/articles/s41586-023-05997-7